Hidden in plain sight

HIDDEN IN PLAIN SIGHT

The untapped potential of written assessment comments

Shiphra Ginsburg

Cover illustration by Douglas Buller

Printing: Datawyse | Universitaire Pers Maastricht

ISBN 978 94 6159 558 4

Hidden in Plain Sight:

The Untapped Potential of Written Assessment Comments

DISSERTATION

to obtain the degree of Doctor at Maastricht University,
on the authority of the Rector Magnificus, Prof. Dr. L.L.G. Soete
in accordance with the decision of the Board of Deans,
to be defended in public
on Thursday, 1st of September, 2016, at 10:00 hours

by

Shiphra Ginsburg

Supervisors

Prof. Dr. Cees van der Vleuten
Prof. Dr. Kevin Eva, University of British Columbia

Co-supervisor

Prof. Dr. Lorelei Lingard, Western University, Canada

Assessment Committee

Prof. Dr. Erik Driessen (chairman)
Dr. Marjan Govaerts
Prof. Dr. Kiki Lombarts, AMC Amsterdam
Prof. Dr. Fedde Scheele, VUMC Amsterdam
Dr. Pim Teunissen

CONTENTS

CHAPTER 1
Introduction

Subjectivity in assessment is gaining increasing respect in the medical education community. The overall goal of this proposed research program is to view subjectivity through the lens of language – what we say and how we say it can provide a window through which we can start to see how clinical supervisors construct opinions and judgments about their learners. Analyzing the language assessors use can deepen our understanding of how they conceptualize competence and performance. Learning how others interpret that language can provide necessary evidence to support the validity of using narrative comments in a way that is credible and defensible.

Assessment in medical education serves many purposes, perhaps most importantly the need to ensure that graduating doctors have "the knowledge and skill required to provide safe and effective care" of their patients.[1]To this end, research over the last several decades has focused on developing instruments aimed at achieving ever higher degrees of objectivity and standardization, such as the checklists commonly used on objective structured clinical exams (OSCEs), while criticizing those that are "too subjective", such as long-form oral exams.[2] Despite these developments, assessment remains a challenge for educators, as too great a focus on objectivity may have led us to preferentially measure what is easy (such as medical expert knowledge) with insufficient attention paid to "those areas for which assessment approaches are more elusive,"[3] such as communication skills or professionalism. This concern is growing in parallel with the advent of competency-based education and assessment models, which require us to "embrace and encompass" a wider variety of perspectives in assessment, such as "fairness, individual needs, safety, reliability … and responsiveness to particular societal and community needs".[3] Thus we need to assess ever more, inter-related competencies that are not particularly well suited to traditional objective measures.

Despite the dominance of the psychometric discourse over the past decades, there is a growing sense that current assessment models are becoming "increasingly threadbare in light of our emerging understanding of the nature of medical practice and of the assessment of medical practitioners".[3] To meet this challenge some have suggested that it's time to reconsider the dominance of the psychometric discourse and to begin to embrace "discourses anchored in subjectivity".[4]

The case for subjectivity:
In an intriguing large-scale study Crossley et al showed that aligning response scales on evaluations with "the priorities of clinician assessors" significantly reduced assessor disagreement and improved reliability of 3 different assessment tools.[5] The authors further noted that each discipline likely has a different conception of what is important in assessing its trainees so consequently the instrument used by each discipline should be unique. What is important, they argue, is the "reality map", or "cognitive structure" of the assessors within each discipline.[6] This suggests that, rather than attempting to make assessments "objective," instead we should strive to better understand influences on (and information derived from) subjectivity such that the added value subjectivity might provide can be effectively built into meaningful assessments. Hodges' work on OSCEs suggested this many years ago by demonstrating that global ratings captured expertise better than did a checklist approach.[7] He has recently argued that if we con-

sider evaluation as being similar to clinical diagnosis, we can quickly see useful parallels – there is great value in the pattern recognition that experts use for diagnosis when compared to the checklist approaches more common to novices.[4] Indeed, the "entrustability" movement relies on supervisors to know when they can trust trainees to perform a given task or role independently.[8] These data together suggest that evaluators' gestalt or gut reactions can have great value, and that we need to tap into what supervisors actually perceive as important in terms of development of competence or expertise.

One way to tap into how supervisors think is to explore the language they use in assessment. As Lingard has succinctly summarized from Kenneth Burke, "language both reflects and shapes our reality".[9,10] How we use language to express ourselves – both what we say and how we say it – can thus provide insights into what and how we think about our trainees. Any educator who has read through stacks of trainee assessments would attest to the fact that we tend to use particular words and phrases repeatedly and in ways that might be different from how they are used in lay communication. Anecdotally, one research student of mine, on reading his first set of assessments, was struck by how pleasant everyone is to work with – a pleasure to work with, a pleasure to have on the team – noting that these phrases seemed odd to him, not being commonly used in his context. What is unknown is why we do we do this. These notions sparked an interest in studying the language assessors use as a way to try to understand what and how they think about their trainees and what they are trying to convey to others.

The case for the utility of analyzing narrative comments in assessment:
In one early study, faculty who were blinded to medical students' numeric scores evaluated the narrative comments on students' assessment forms.[11] They identified significantly more students as being in difficulty when compared to the students' tutors who were the ones who completed the forms. In other words, negative or problematic comments were present on students' assessments, but they were not reflected in the assigned scores. The authors concluded that analyzing the comments on evaluation forms rather than simply looking at the numbers would be beneficial to students' progress. In a similar study in the postgraduate setting, Durning et al reported an analysis of over 1500 Internal Medicine (IM) residents' evaluation forms collected over a 10 year period.[12] Although most comments were concordant with the numeric scores assigned, a significant minority of the comments were coded as more negative than the scores would suggest, particularly in the domains of attitude and clinical skills. Again, the authors concluded that the comments can reveal areas requiring attention and remediation that would otherwise have been missed if focus is constrained to the numeric scores. One somewhat contradictory study involving peer-assessment of 302 practicing physicians in the UK found that most of the peers' comments were "aligned with" the scores assigned so the authors concluded that "there was little benefit in routinely analysing narrative comments for the purposes of revalidation".[13] Their efforts, however, were focused on finding negative comments that contradicted a positive scale rating because for purposes of revalidation it was important to determine whether relying solely on scores would be appropriate. They did note that approxi-

mately 8% of forms had at least one "adverse" statement, suggesting that the comments could still be a source of useful feedback to physicians.

Insight into why this might be so can be gained from a study of In-training Evaluation Reports or ITERs that we conducted prior to carrying out the research included in this dissertation. Since residents obtain ITERs after each clinical rotation (10-12 per year) and each ITER contains numeric scores as well as boxes for free-text comments they could serve as a source of narrative commentary suitable for studies similar to those above. (See below for more details regarding our ITER form and assessment system). In order to explore this potential we conducted a qualitative analysis of comments using constructivist grounded theory approach, which allows the researcher to take an open, inductive approach to coding while being mindful of findings from similar research.[14] The database for this study contained the narrative comments from 1770 ITER forms collected from 180 residents in our internal medicine (IM) program. Many of the comments were found to map onto one or more competency domains, such as knowledge base, often linking to several overlapping competencies at once. Of interest, many comments were completely unrelated to competencies, such as the resident's personality, disposition or impact on staff. Supervisors frequently noted a resident's apparent trajectory – how they appeared in relation to expected norms, whether or how they improved over time, and how they were predicted to perform in the future. These constructs were not otherwise captured on the numeric portion of the ITER, suggesting that ITERs, as structured, were not optimally capturing preceptors' integrated, holistic impressions and that perhaps we would gain a more complete picture of residents' performance by analyzing the language more closely.

These studies all highlight that discordance can be found between numeric ratings and comments, yet it is not totally clear why this might arise. It is possible that the ratings and the comments each pick up different aspects of "competence" and thus might be expected to be different in some way, but that does not explain why they might be discordant, and in disagreement with one another. It is also possible that the process of translating from an observation to an impression and then into a series of numeric scores can be problematic, with a resulting loss of information and authenticity.[15] The literature suggests that grade-inflation and a "failure to fail" culture may contribute to this phenomenon.[16] These issues are important to study and tease out as we try to gain a better understanding of the potential contributions of language analysis to assessment practice.

Considerations of how language might be analyzed:
Thus, evidence suggests that the numeric ratings found on ITER forms may not be well suited to capturing supervisors' complete impressions, whereas conducting in-depth qualitative analyses of comments can provide useful information that is potentially more educationally relevant; however, such analyses can be very labour-intensive. In response, some researchers have investigated the utility of "scoring" qualitative comments according to their positive/negative polarity (what some have called valence). In one study of the evaluation of professionalism, for example, investigators found that the number of "positive" written comments on a student's evaluation form correlated

with students' numerical professionalism score while the number of "negative" comments, although uncommon, had a negative correlation.[17] It is worth noting that the number of "equivocal" comments also had a significantly negative correlation; that is, they behaved exactly like negative comments. This suggests that it might be fruitful to look for these equivocal or lukewarm comments as they may indicate more significant issues for the student. However, not only is it awkward to use the numerical rating as an outcome when the argument for the value of narrative comments is that they contain better information, but the actual process of scoring comments for polarity can be problematic. For example, in Frohna's study comments were coded as "equivocal" if they were neither wholly positive nor wholly negative, and as a result these comments often contained both positive and negative phrases. The authors were clear that these comments were not, in fact, "neutral". Canavan et al parsed multiple or complex statements into "feedback phrases" with a single point of focus, so that each could be scored as either positive or negative.[18] However this did not entirely solve the problem either as polarity could not be assessed for up to 20% of comments.

Further, it is interesting to consider that the studies described above all attempted to reduce language to smaller pieces that could then be coded in some way. Yet there is some research suggesting that it might be important to analyze complete sentences or longer passages rather than parsing them into phrases. For example, in Frohna's study, if those "equivocal" phrases were split into the wholly positive and wholly negative components the authors would have lost their ability to determine the overall effect of these complex phrases. In addition, Bogo et al. found that when faculty social workers described students' problem areas they often used "'but' statements." For example, stating that a student was liked by the staff *but* was too casual and had poor boundaries with clients.[19] Similarly, in an interview study by my research team, we found that experienced faculty attendings would often dismiss areas of strength in residents who were problematic, whereas they would discount or excuse deficiencies in residents who they thought were outstanding.[15] For the problematic residents, these areas of strength were often phrased as "but statements", as faculty often began their descriptions with a positive comment followed by a "but…". These phrases may be important indicators of faculty's opinions, but will not be captured well either by coding the statement as "equivocal" or by segmenting the sentence into two separate, equally weighted "feedback phrases".

Further, in none of the described studies was the *degree* of positivity or negativity of comments noted. That is, comments were either positive, negative, equivocal or neutral, but no effort was made to determine *how* positive or *how* negative. The most likely reason for this (personal communication with several authors) is that it can be difficult to agree on the degree to which a comment is positive or negative. It might be easy to agree that adjectives describing knowledge base such as excellent, outstanding or superb might reflect the same underlying meaning, but it may be more problematic to quantify the polarity of phrases like decent, solid, adequate, or appropriate. These are all "positive" but clearly less so and they can even be interpreted negatively depending on context (e.g., by providing lukewarm praise).[20] In fact, recent research has shown that such adjectives may have different connotations depending on which do-

mains of performance are being assessed.[21] It is unclear what effect this might have on the relationships noted thus far between comments and scores. These more nuanced interpretations may be obscured by simplified coding schema.

In comparison to research based on analyzing short phrases, some researchers have explored the use of more "holistic, realistic vignettes" of trainees as a basis for assessment and found that using these vignettes, or profiles, led to improved supervisor discrimination of learners' performance in social work.[22] Supervisors were also more easily able to identify learners in difficulty. My research team built on this work by asking faculty to categorize and rank-order holistic, narrative profiles of IM residents.[23] For this purpose we used one-page profiles that were created based on data from interviews with faculty in which they described outstanding and problematic residents with whom they had worked.[15] We found that faculty could rank-order residents based on the comments alone with impressive reliability (single-rater intraclass correlations were in the range of 0.8-0.9), leading us to some of the study design choices we made (e.g., asking judges to evaluate residents using an aggregation of comments) outlined in the chapters that follow.

These studies taken together illustrate the potential value in analyzing language in assessment comments but also highlight difficulties and problems that can arise when trying to reduce language to smaller, digestible pieces that can then be counted and quantified: meaning can get lost along the way. Analyzing language in longer, more naturalistic passages might yield a more complete and nuanced understanding.

In summary, then, in the medical education research community there is an increasing recognition of the value inherent in a subjective, collective approach to assessment. Viewing subjectivity through the lens of language can provide a window through which we can start to see how clinical supervisors construct their opinions and judgments about learners. Evidence suggests that the narrative comments on assessments can provide important additional sources of assessment data, above and beyond numeric scores, perhaps even improving the ability of ITERs to identify residents in difficulty. Yet in practice comments are rarely used in a systematic way, in part due to the labour involved and also because there is uncertainty about what to do if the scores and comments are discordant. Finally, the education community in general has, until recently, been reluctant to embrace "subjective" means of assessment, greatly preferring quantitative and more apparently "objective" methods.[24,25]

With these thoughts in mind, several overarching goals took shape as a basis for this dissertation. Given the discussion on the untapped potential of ITER comments in assessment one of my major goals was to determine if narrative comments could be used for assessment in a way that is reliable, credible and has validity for its intended purpose. Findings from these studies could have broad, immediate impact and would be timely given the challenges posed by the advent of Competency-Based Medical Education (CBME). A second goal was to gain a much deeper and more nuanced understanding of the language attendings use when assessing their residents in IM. The research summarized above suggests that assessment language is often non-literal and

context-bound and thus it might be difficult to interpret. Given the educational potential of these assessments I wanted to study how comments are constructed, why attendings write this way, what it means to others and what that might say about assessment as a whole. Specifically, through a series of linked research studies I will attempt to address each of the following research questions.

> Research Question #1
> Can narrative comments on ITERs be used reliably to discriminate between residents in Internal Medicine?
> This question is addressed in chapters 2, 4 and 6.
>
> Research Question #2
> How do readers make sense of narrative comments?
> This is addressed largely in chapters 3 and 6, and in a complementary way with data presented in chapter 5.
>
> Research Question #3
> Can narrative comment analysis be used as a feasible approach to assessment?
> This is largely addressed in chapter 4, with supporting data presented in chapters 2 and 6)
>
> Research Question #4
> Why do clinical supervisors write the way they do?
> This is the focus of chapter 5, with supporting information presented in chapters 3 and 6.
>
> Research Question #5
> Do residents themselves understand what is written in ITER comments? What educational value do they perceive from the comments?
> This is addressed in chapter 6.

A brief summary of specific chapters follows:

Chapter 2
This chapter represents results of the first study, which was designed to explore the reliability and predictive ability of scores and narrative comments on ITERs from postgraduate years 1-3 (PGY1-3) in our Internal Medicine (IM) program. This was a mixed-methods study involving 24 participants using a database of ITERs collected from all PGY1-3 residents graduating in 2011. Findings from this study largely provide evidence in response to RQ1, with some supporting evidence for RQ3.

Chapter 3
To understand how faculty participants interpreted and made sense of the ITER comments they rank-ordered in the experimental phase described in chapter 2, interviews

were conducted with the 24 faculty participants and analyzed using constructivist grounded theory. The framework developed in this study provides evidence in response to RQ2, and led to the development of RQs 4 and 5.

Chapter 4
The study reported in chapter 4 builds on findings from Chapters 2 and 3 in two ways: by examining the potential advantage of being an "insider" when judging assessment comments and by exploring the effects on reliability when less data are available.

The first objective was to determine whether findings would be replicable using a different data set and attendings from outside of our system. The participants reading and interpreting the comments in Chapters 2 and 3 were attending physicians in the same program through which the comments were generated, which could give them an insider advantage. That is, they may be familiar with the way performance is described for these residents and could be influenced by certain elements of language and structure that might be particular to our assessment culture. For this reason we recruited participants who were IM attending faculty from institutions across Canada but external to the University of Toronto. This step also allowed for validation of findings in a different cohort (one that graduated in 2010). The second objective was to address the question of whether the reliabilities seen in chapter 2 could be attributed (at least in part) to the volume of comments assessed. That is, participants had been given an entire year's worth of comments for each PGY1 they rank-ordered, and even ITER scores can gain decent reliability when collected over a year. To know whether the comments have an edge in this regard we also assessed the reliability of rank-ordering by comments alone when only the first three comments of the year were given to participants.

This study involved 24 participants from university IM programs across Canada and is reported in Chapter 4. This paper provides much of the evidence in response to RQ3 and also supporting evidence for RQ1.

Chapter 5
This study was conducted in an attempt to address questions that arose from our earlier studies. In particular, we struggled to reconcile the finding that much of assessment language is considered vague and generic with the finding of high inter-rater reliability when using that language for rank-ordering residents. We wanted to understand how otherwise vague language could also be so meaningful. For this study we turned to the theory of "politeness", situated in the branch of linguistics called pragmatics and conducted an in-depth analysis of a subset of the PGY1 ITER comments from the 2011 cohort. This study provides evidence in response to RQ4 and sets the stage for future research.

Chapter 6
This chapter represents the final study conducted, in response to questions that arose throughout the previous 4 chapters. Importantly, it remained to be seen whether residents could effectively interpret the narrative comments and whether their interpreta-

tion matches the faculty's. In this study we replicated the methodology first seen in Chapters 2 and 4 but with 12 PGY2 residents from our own IM program as participants. We used a mixed-methods approach involving a quantitative analysis of reliability as well as a qualitative analysis of interviews using constructivist grounded theory. This study provides direct evidence in response to RQ5 and also has bearing on all other RQs posed.

Chapter 7
In this chapter findings from all studies will be integrated, synthesized and discussed as a body of work, drawing on four major themes that were identified during collective interpretation. Discussion of the advantages and constraints of the methodologic approach taken will also be considered. The chapter concludes with a consideration of implications for practice and for future research,

Research setting and context
As all of the studies in this dissertation used ITER data from the University of Toronto I will provide an overview of the educational structure of our IM residency as well as our ITER forms.

The University of Toronto has Canada's largest IM program, with 55-60 residents per year. The first three postgraduate years (PGY1-3) are considered "core" years, following which residents must do either an additional 2-year subspecialty program (such as Oncology or Nephrology) or do a final year of General Internal Medicine (GIM) before they can enter independent practice. The subspecialty match during PGY3 is competitive and uses a national matching service. In our program each resident rotates through one month blocks throughout the year (e.g., GIM – also known as the CTU for clinical teaching unit), Cardiology, Nephrology, etc) with each rotation generating a single ITER that is filled out by the attending physician (with exceptions, as will be noted). There is no set order to the rotations but each resident does four months of GIM in their PGY1 year. Many residents take research blocks and some are allowed external elective time. That combined with the unfortunate fact that ITERs are occasionally not submitted means that residents receive an average of about 10 ITERs per year.

At the time of the study our ITER form consisted of 18 items related to the seven CanMEDs competencies,[26] which is our guiding framework for assessment in Canada. There were 4 items on the Medical Expert role, 3 each for the Scholar, Health Advocate and Professional roles and 2 each for Communicator, Collaborator and Manager. There was also an item at the end for an "overall rating" which has no further description. All items are scored on a scale of 1-5 as follows: 1 = unsatisfactory, 2 = needs improvement, 3 = meets expectations, 4 = exceeds expectations, and 5 = outstanding. Instructions at the top of the form state that fewer than 5% of residents should receive a 5, 20-40% should receive a 4 and the majority should receive a 3. Very few should fall into the 1 or 2 categories.

At the end of the form there is a single box for general comments with the instruc-

tions: “Provide a general impression of the trainee’s development during this rotation, including general competence, motivation and consultant skills. Please emphasize strengths and areas that require improvement. If 'exceeds expectations', 'needs improvement', and/or unsatisfactory' ratings have been assigned, provide the supporting comments in this space.”

Our ITERs are electronic and are managed locally through a Research Officer (RO) in the Department of Medicine. The RO provided us with all of the raw data for these studies and was responsible for anonymizing the narrative data and providing unique identifiers so that residents could be tracked through their rotations.

Methods and Methodology

As can be appreciated from the brief descriptions of each study, multiple methods and methodologic approaches were taken throughout this work. Together, these studies are best thought of not necessarily as linear or sequential but rather as a triangulation of methodologies and approaches, each contributing a unique perspective. No one methodology or study design would be adequate to answer the qualitative and quantitative questions posed above. It therefore became apparent that a mixed-methods approach would be most appropriate; in particular we used a multiphase, mixed-methods design as described by Creswell and Clark.[27]

Mixed-methods research has been gaining traction in medical education and has been described as being particularly beneficial when studying “new questions or complex initiatives and interactions”.[28] For this thesis I was influenced by the work of John Creswell, an educational psychologist and prominent advocate of mixed methods research in education and health sciences.[29] In a paper on “Best Practices in Mixed Methods Research”, Creswell defines mixed methods research as an approach or methodology that embraces a number of features, including a focus on research questions that require real-life contextual understanding, multiple perspectives and acknowledgment of cultural influences; using multiple methods (such as experiments and interviews); employing rigorous quantitative techniques (to assess magnitudes and frequencies) and qualitative techniques (to explore the meaning and understanding of constructs); and intentionally integrating or combining these methods to draw on the strengths of each.[27]

The research questions posed above are ideally suited to mixed-methods exploration. They are grounded in a desire to understand a real-life phenomenon in context from multiple perspectives. My studies used multiple methods, including experimental design, interviews, and analysis of text, and each was conducted with attention to the appropriate principles of rigour, as will be specified more deliberately in each chapter. Integration is a key feature of mixed-methods research as it provides the opportunity to combine qualitative and quantitative data for interpretation rather than keeping each part separate. In multiphase designs integration is best appreciated at the level of the body of work as a whole, rather than within each study.

In terms of study design, although mixed-methods research is often thought of simply as studies that use both qualitative and quantitative data, there are specific design frameworks that usually underpin an investigation, whether explicitly stated or not. For example, some studies take a sequential approach, with either qualitative or quantitative data collection coming first followed by the other. These studies may use the qualitative data to either explain or explore the quantitative, or vice versa. Some designs are parallel or convergent, with data collection for each taking place side by side and only being integrated later. My study design would be categorized by Creswell and Plano Clark's framework as "multiphase" which involves multiple projects conducted over time and "linked together by a common purpose," with studies designed so that each phase builds on the one before.[27] The choices made for each study were driven by a pragmatic goal: to use "diverse approaches, giving primacy to the importance of the research problem and question, and valuing both objective and subjective knowledge."[27] Following Johnson, my mixed methods research sought to "use a method and philosophy that attempt to fit together the insights provided by qualitative and quantitative research into a workable solution."[30]

The "integration" step is not easily achieved within individual studies and will be illustrated fully in the Discussion (Chapter 7). However, I will highlight here the approach to be taken so that it may be kept in mind when reading each chapter. There are different approaches to attempting to integrate findings, such as using the qualitative piece to explain the quantitative findings or vice versa. Creswell discusses three major approaches: merging, connecting and embedding. Merging involves presenting data together as a means to create integration, with each piece "speaking to" the other pieces so that they may be understood as a whole. In the connecting approach, integration occurs by connecting the analysis from one phase of research to subsequent phases. Embedding achieves integration by having one dataset subsumed in a larger, primary dataset. In terms of the individual studies making up this thesis, some examples can be seen: in Studies 1 and 4 I present merged results; however, the overall approach I took is one of connecting, which best fits with the multiphase study design.

The chapters that follow should therefore be considered as a multiphase, mixed-methods project (Chapters 2-6) with integration based on a connected approach in Chapter 7.

References

1. Eva KW, Bordage G, Campbell C, Galbraith R, Ginsburg S, Holmboe ES, et al. Towards a program of assessment for health professionals: from training into practice. *Adv Heal Sci Educ* 2015;Online early.
2. Eva KW, Hodges BD. Scylla or Charybdis? Can we navigate between objectification and judgement in assessment? *Med Educ* 2012;46(9):914–9.
3. Whitehead CR, Kuper A, Hodges BD, Ellaway R. Conceptual and practical challenges in the assessment of physician competencies. *Med Teach* 2015;37(3):245–51.
4. Hodges BD. Assessment in the post-psychometric era: Learning to love the subjective and collective. *Med Teach* 2013;35(7):564–8.
5. Crossley J, Johnson G, Booth J, Wade W. Good questions, good answers: construct alignment improves the performance of workplace-based assessment scales. *Med Educ* 2011;45(6):560–9.
6. Crossley J, Jolly BC. Making sense of work-based assessment: ask the right questions, in the right way, about the right things, of the right people. *Med Educ* 2012;46(1):28–37.
7. Hodges BD, Regehr G, McNaughton N, Tiberius R, Hanson M. OSCE checklists do not capture increasing levels of expertise. *Acad Med* 1999;74(10):1129–34.
8. Sterkenburg A, Barach P, Kalkman C, Gielen M, ten Cate OT. When do supervising physicians decide to entrust residents with unsupervised tasks? *Acad Med* 2010;85(9):1408–17.
9. Lingard L, Garwood K, Szauter K, Stern DT. The rhetoric of rationalization: How students grapple with professional dilemmas. *Acad Med* 2001;76(10 suppl):S45–7.
10. Ginsburg S, Lingard L. Using Reflection and Rhetoric to Understand Professional Behaviors. In: Stern DT, editor. Measuring Medical Professionalism. New York, New York: Oxford University Press; 2006. page 195–212.
11. Cohen G, Blumberg P, Ryan N, Sullivan P. Do final grades reflect written qualitative evaluations of student performance? *Teach Learn Med* 1993;5(1):10–5.
12. Durning SJ, Hanson J, Gilliland W, McManigle JM, Waechter D, Pangaro LN. Using Qualitative Data From a Program Director's Evaluation Form as an Outcome Measurement for Medical School. *Mil Med* 2010;175:448–52.
13. Richards SH, Campbell JL, Walshaw E, Dickens A, Greco M. A multi-method analysis of free-text comments from the UK General Medical Council Colleague Questionnaires. *Med Educ* 2009;43(8):757–66.
14. Ginsburg S, Gold W, Cavalcanti RB, Kurabi B, McDonald-Blumer H. Competencies "Plus": The nature of written comments on internal medicine residents' evaluation forms. *Acad Med* 2011;86(10 Suppl):s30–4.
15. Ginsburg S, McIlroy J, Oulanova O, Eva KW, Regehr G. Toward authentic clinical evaluation: pitfalls in the pursuit of competency. *Acad Med* 2010;85(5):780–6.
16. Dudek NL, Marks MB, Regehr G. Failure to fail: the perspectives of clinical supervisors. *Acad Med* 2005;80(10 Suppl):S84–7.
17. Frohna A, Stern DT. The nature of qualitative comments in evaluating

professionalism. *Med Educ* 2005;39(8):763–8.
18. Canavan C, Holtman MC, Richmond M, Katsufrakis PJ. The quality of written comments on professional behaviors in a developmental multisource feedback program. *Acad Med* 2010;85(10 Suppl):S106–9.
19. Bogo M, Regehr C, Woodford M, Hughes J, Power R, Regehr G. Beyond Competencies: Field Instructors' Descriptions of Student Performance. *J Soc Work Educ* 2006;42(3):579–93.
20. Kiefer CS, Colletti JE, Bellolio MF, Hess EP, Woolridge DP, Thomas KB, et al. The "Good" Dean's Letter. *Acad Med* 2010;85(11):1705–8.
21. Kan Ma H, Min C, Neville A, Eva KW. How good is good? Students and assessors' perceptions of qualitative markers of performance . *Teach Learn Med* 2013;25(1):15–23.
22. Regehr G, Bogo M, Regehr C, Power R. Can We Build a Better Mousetrap? Improving the Measures of Practice Performance in the Field Practicum. *J Soc Work Educ* 2007;43(2):327–43.
23. Regehr G, Ginsburg S, Herold J, Hatala R, Eva KW, Oulanova O. Using "Standardized Narratives" to Explore New Ways to Represent Faculty Opinions of Resident Performance. *Acad Med* 2012;87(4):419–27.
24. Norman GR, van der Vleuten CPM, Graaff E. Pitfalls in the pursuit of objectivity: issues of validity, efficiency and acceptability. *Med Educ* 1991;25(2):119–26.
25. van der Vleuten CPM, Norman GR, Graaff E. Pitfalls in the pursuit of objectivity: issues of reliability. *Med Educ* 1991;25(2):110–8.
26. Frank JR. The CanMEDS 2005 Physician Competency Framework. Better standards. Better physicians. Better Care. Ottawa: The Royal College of Physicians and Surgeons of Canada; 2005.
27. Creswell JW, Klassen AC, Plano VL, Smith KC. Best practices for mixed methods research in the health sciences. A report commissioned by the Office of Behavioural and Social Sciences Research. 2011.
28. Schifferdecker KE, Reed VA. Using mixed methods research in medical education: basic guidelines for researchers. *Med Educ* 2009;43(7):637–44.
29. Creswell J, Clark V. Designing and conducting mixed-methods research. SAGE; 2011.
30. Johnson RB, Onwuegbuzie AJ. Mixed Methods Research: A Research Paradigm Whose Time Has Come. *Educ Res* 2004;33(7):14–26.

CHAPTER 2

Do In-Training Evaluation Reports Deserve Their Bad Reputations? A Study of the Reliability and Predictive Ability of ITER Scores and Narrative Comments

Published as: Ginsburg S, Eva KW, Regehr G. Do in-training evaluation reports deserve their bad reputations? A study of the reliability and predictive ability of ITER scores and narrative comments. *Acad Med*. 2013;88(10):1539-1544.

Abstract

Purpose

Although scores on in-training evaluation reports (ITERs) are often criticized for poor reliability and validity, ITER comments may yield valuable information. The authors assessed across-rotation reliability of ITER scores in one internal medicine (IM) program; ability of ITER scores and comments to predict post-graduate year three (PGY3) performance; and reliability and incremental predictive validity of attendings' analysis of written comments.

Method

Numeric and narrative data from the first two years of ITERs for a single cohort of residents at the University of Toronto Faculty of Medicine (2009-2011) were assessed for reliability and predictive validity of third-year performance. Twenty-four faculty attendings rank-ordered comments (without scores) such that each resident was ranked by 3 faculty members. Mean ITER scores and comment rankings were submitted to regression analyses with the dependent variables being PGY3 ITER scores and program director's rankings.

Results

Reliabilities of ITER scores across 9 rotations for 63 residents were .53 for both PGY1 and PGY2. Inter-rater reliabilities across 3 attendings' rankings were .83 for PGY1 and .79 for PGY2. There were strong correlations between ITER scores and comments within each year (.72 and .70). Regressions revealed that PGY1 and PGY2 ITER scores collectively explained 25% of variance in PGY3 scores and 46% of variance in PGY3 rankings. Comment rankings did not improve predictions.

Conclusions

ITER scores across multiple rotations showed decent reliability and predictive validity. Comment ranks did not add to the predictive ability but correlation analyses suggest that trainee performance can be measured through these comments.

Evaluating residents' clinical competence is challenging for educators. In-training evaluation report (ITER) scores are used by most residency programs to perform this task despite suggestions that they are not reliable in practice and suffer from validity issues.[1] These evaluations are often done without proper training, are not based on enough direct observation, and suffer from recall bias, among other issues. Further, the literature on "failure to fail" suggests that clinical teachers are reluctant to provide honest yet critical feedback due to multiple cognitive, social, and political biases.[2] As a result, calls have been made to reduce our reliance on these types of assessments.[3]

In contrast to the issues with the numerical data that ITERs generate, some research suggests that the comments written on these forms may be a better source of information. In one study faculty participants who analyzed only the written comments on evaluation forms were much more likely to identify problem students (and at earlier time points) than when using the scores alone.[4] Thus there was a discordance between what was written and the grades assigned. More recent studies of residents and practicing physicians have also found a significant number of comments that were discordant with scores, generally indicating a lower performance level than the scores would suggest.[5,6] Importantly, one study found that "equivocal" comments, neither wholly positive or negative, were negatively correlated with scores and in fact behaved just like negative comments.[7] In other qualitative work, faculty's written comments on residents' evaluations were found to offer additional, interpretive information that did not always map easily onto standard competencies.[8] Thus, it seems there is richness in the narrative comments that the scores do not capture, perhaps due to difficulty in translating the holistic nature of one's perceptions or concerns into simple numerical ratings. Systematically reading and scoring comments is problematic, however, in that it is not always easy to decide whether a comment is positive, negative or somewhere in-between.

As analyzing written language is so labor intensive, several software programs have been developed that can automate this task; however, before any of these programs can be evaluated for use on ITER comments certain assumptions require testing including whether or not narrative comments from ITER evaluations can be reliably and meaningfully assessed. To evaluate these assumptions we conducted a study to predict the performance of one cohort of third-year residents (PGY3) based on ITER scores and comments from their first and second postgraduate years (PGY1 and PGY2). Our hypotheses were as follows: ITER scores are unreliable across rotations, as commonly reported; as a result, ITER scores from PGY1 and PGY2 are poor predictors of performance in PGY3 as measured by both ITER scores and program directors' (PD) rankings; and in line with the research synthesized in the previous paragraph, faculty attendings' analysis of narrative comments from the ITERs can be reliable and will be more predictive than ITER scores.

Method

Our dataset was a compilation of all ITERs from a single cohort of residents collected across PGY 1-3 in a core internal medicine (IM) training program at the University of Toronto Faculty of Medicine, for the academic years 2009-2011. The data were anonymized and divided by academic year. In this program, residents are expected to receive a single ITER at the end of every one-month rotation.

The dependent variable of interest was performance during PGY3, which we assessed in two ways. The first outcome variable was the mean ITER score for each resident across all ITERs collected during their PGY3 year. The second outcome variable was created by inviting PDs from the major teaching sites within our program to rank-order their most recent group of PGY3s. Three of four site PDs are responsible for 12-15 residents each, and one has the responsibility for two hospitals combined with a total of 29 residents. To reduce recall bias as much as possible we conducted this ranking task between July and August for all PGY3s who graduated in June of that year. PDs were shown a list of their most recent PGY3s and were asked to rank-order them from highest to lowest based on whatever relevant knowledge they had and to categorize them according to a previously developed framework: A = outstanding, excellent, exemplary; B = solid, safe, may need some fine tuning; C = borderline, bare minimum, remediable; D = unsafe, unacceptable, multiple deficits.[9] They were allowed to review any notes or prior ITERs they had for each trainee. Because each site had different numbers of trainees we scaled the rankings to enable meaningful collation of rankings across sites. That is, the highest ranked resident from each site received a score of 1, the lowest ranked resident from each site received a score of 0, and those in between received scores corresponding with their percentile. Thus we had 2 scores for each PGY3: the mean ITER score (ranging from 1-5) and a PD rank-order score within each of the four teaching sites (ranging from 0 to 1).

ITER data

Our ITERs consist of 19 items related to the seven CanMEDs[10] competencies. There are four items on the Medical Expert role, two each for the Communicator, Collaborator, and Manager roles, and three each for the Scholar, Health Advocate, and Professional roles. There is also an overall rating. All items are scored on a 5-point scale as follows: 1 = unsatisfactory, 2 = needs improvement, 3 = meets expectations, 4 = exceeds expectations, 5 = outstanding. We calculated the mean of all 19 items plus the overall rating for each resident for each rotation to generate a score ranging from 1 to 5 for each ITER.

Narrative data

Each ITER form has a single box for general comments with instructions to "Provide a general impression of the trainee's development during this rotation, including general competence, motivation and consultant skills. Please emphasize strengths and areas that require improvement." We cut and pasted the comments from each ITER and combined them for each resident separately for PGY1 and PGY2. Thus, each resident had two documents of comments, 1-2 pages in length, one for each year, with no numerical data attached. For each year separately, these documents were randomly assigned to 12 sets of 15-16 documents, such that each resident's comments appeared in three different sets, no two sets were alike, and no identifying information (regarding evaluator or resident) was included.

Narrative ranking procedure

Twenty-four faculty attendings with at least 2 years' experience evaluating residents on inpatient units in IM were recruited by email to participate. A research assistant conducted one-on-one sessions, during which each participant was given up to an hour to read through a distinct package of residents' comments and to rank-order them from 1-15 or 16. The research assistant conducted a debriefing interview with each participant upon completion of the exercise.

Analysis

To determine the internal structure of the scale we conducted a factor analysis of all fully complete ITERs in the dataset.

Reliability of the ITER scores for each resident was calculated across ITERs within each year using Shrout and Fleiss Case 1[11] in an unbalanced design to allow for the inclusion of residents with different numbers of ratings (each resident had between 1 and 13 ITERs with an average of 9.63). To allow consistent comparisons across groups, the inter-ITER reliability is reported with k (the number of observations averaged across) set to 9 (the approximate average number of ITERs per year across residents in our data set).

Reliability of the comment-ranking procedure across three raters was assessed by using Shrout and Fleiss Case 1 (with raters nested within residents). Thus we calculated 3-rater reliability (k = 3) for the ranking of each resident's comments.

We also calculated the correlation between the mean PGY3 ITER scores and the PD's rankings.

To assess predictive ability of the ITER scores and comment rank scores we conducted regression analyses with PGY3 performance as our dependent measures (separately for

ITER scores and PD rankings). For each regression, we first used PGY1 and PGY2 ITER scores as independent variables, then added the comment rank scores for both PGY 1 and PGY2 to assess for incremental predictive ability (increase in R^2).

Ethics approval was obtained from the University of Toronto's Office of Research Ethics.

Results

Of a total of 75 possible residents, 63 (84%) had sufficient data to include in the analysis, but only 59 received rankings from the PGY3 site PDs. Thus, for all analyses presented, our sample size was 63 residents except for the regression predicting PGY3 PD rankings, where the sample was 59.

Analysis of the dimensions of the ITER scales

The factor analysis utilized 903 fully completed forms accrued across resident, rotation, and year. The unrotated factor solution revealed two factors, with the first accounting for 66.0% of the variance and every item loading more or less equally from a low of .69 for basic science knowledge to a high of .85 for the overall rating at the end of the scale. The second factor accounted for an additional 5.4% of the variance. A varimax rotation revealed two major dimensions, which can be roughly translated as "knowledge/clinical skills" and "interpersonal skills" (see Table 1). Despite having seven competency roles represented on our ITERs we did not see obvious clustering into these roles. Interestingly, the overall or global rating on the ITER was the highest loading variable on the unrotated general factor and was nearly equally weighted between the two factors in the rotated solution (with a slight bias toward knowledge/skills), suggesting it was truly an overall rating of the resident.

Predictor variables

The reliability of a single ITER score was 0.11 for both the PGY1 and PGY2 data. However, the reliability for the average of 9 rotations was moderately high at 0.53 for both years. For the ranking of the ITER comments the reliability for 3 raters per resident was 0.83 for PGY1 and 0.79 for PGY2.

Table 1. Results of Factor Analysis ((Unrotated and Varimax) for All Items on 903 In-Training Evaluation Reports for First-, Second-, and Third-Year Residents, University of Toronto Faculty of Medicine, 2009-2011*

		Unrotated factor analysis		Varimax rotation	
Item on in-training evaluation report	CanMEDS role†	Factor 1	Factor 2	Knowledge or clinical skills	Interpersonal skills
Q15 Demonstrates an effective evidence based approach to practice	Sch	0.784	0.341	0.794	0.317
Q02 Acquires and exhibits appropriate clinical knowledge	ME	0.784	0.33	0.786	0.325
Q01 Demonstrates basic science knowledge	ME	0.688	0.392	0.763	0.213
Q03 Demonstrates effective clinical decision making skills	ME	0.807	0.209	0.716	0.427
Q11 Demonstrates attentiveness to preventive measures	HA	0.834	0.153	0.696	0.485
Q20 Overall rating (or "global")		0.852	0.137	0.696	0.510
Q04 Demonstrates appropriate procedure skills	ME	0.794	0.169	0.678	0.445
Q10 Utilizes healthcare resources appropriately	Mgr	0.846	0.084	0.655	0.542
Q16 Facilitates the education of patients, students, others	Sch	0.838	0.083	0.649	0.538
Q12 Recognizes important determinants of health	HA	0.839	0.077	0.645	0.542
Q14 Attends and contributes to rounds, learning events	Sch	0.809	0.059	0.611	0.534
Q09 Manages time effectively	Mgr	0.798	0.031	0.583	0.546
Q06 Provides clear, concise oral and written reports	Comm	0.812	−0.09	0.507	0.641
Q13 Advocates effectively on behalf of patients	HA	0.835	−0.17	0.467	0.713
Q17 Demonstrates appropriate self-awareness, insight, of abilities	Prof	0.828	−0.181	0.454	0.716
Q05 Communicates with and counsels patients and families effectively	Comm	0.806	−0.232	0.402	0.736
Q19 Demonstrates respect for diversity, maintains appropriate boundaries	Prof	0.844	−0.300	0.380	0.811
Q07 Works effectively in a multidisciplinary environment	Coll	0.822	−0.32	0.351	0.809
Q18 Demonstrates reliability and responsibility	Prof	0.789	−0.3	0.342	0.772
Q08 Maintains respectful relationships in the workplace	Coll	0.825	−0.383	0.308	0.856

*Items are sorted by factor loadings in the varimax rotation.
†Sch indicates scholar; ME, medical expert; HA, health advocate; Mgr, manager; Comm, communicator; Prof, professional; Coll, collaborator.

Outcome measures

The reliability of the PGY3 ITER scores was 0.14 for a single ITER and 0.59 for the average of 9 rotations. Figure 1 shows the correlation (r = .63) between the average ITER scores and the PD rank-orderings (scaled from 0 to 1). The grey dots represent residents that PDs placed in category C (borderline performance) and the single black dot represents one resident that a PD placed in category D (multiple deficits, unsafe, unacceptable).

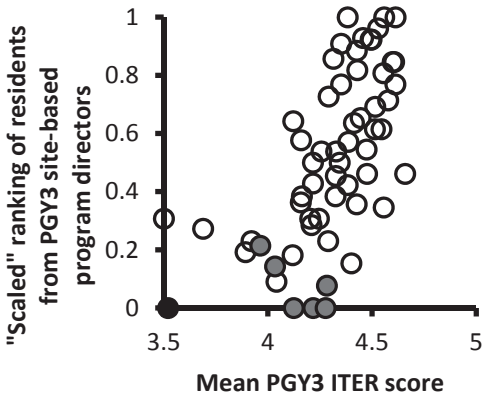

Figure 1. Correlation between the average in-training evaluation report (ITER) scores and the site program directors' (PD) rank-order scores (scaled from 0 = worst to 1 = best) for all postgraduate year three (PGY3) residents, University of Toronto Faculty of Medicine, 2009-1011. Gray dots represent residents placed by their PDs in category C (borderline, bare minimum, remediable); the single black dot represents a resident placed in category D (multiple deficits, unsafe, unacceptable).

Predicting PGY3 assessments

Table 2 shows the correlations between the PGY1 and PGY2 average ITER scores and the comment rank-order scores compared with the two PGY3 outcomes (ITER scores and PD rank-ordering). There were strong correlations between ITER scores and comment scores within each of PGY1 and PGY2 (.72 and .70), suggesting that much of the information gleaned from reading and ranking the comments was already captured in the numbers, with a shared variance of about 50%. Correlations between both predictor scores and the PGY3 ITERs were moderate (.44 and .44 for PGY1 and .44 and .33 for PGY2). These correlations were higher for the PGY3 PD rank-order score (.58 and .53 for PGY1 and .60 and .62 for PGY2).

Table 2. Correlations Between the Postgraduate Year One (PGY1) and PGY2 In-Training Evaluation Report (ITER) Scores and Comment Rank-Order Scores, Compared with PGY3 ITER scores and Program Directors' (PD) Rank-Ordering, University of Toronto Faculty of Medicine, 2009-2011

Predictor	PGY1 average ITER score	PGY1 comment rank-order score	PGY2 average ITER score	PGY2 comment rank-order score
PGY1 comment rank-order score	.72	--	--	--
PGY2 average ITER score	.56	.52	--	--
PGY2 comment rank-order score	.53	.59	.70	--
PGY3 average ITER score (n=63)	.44	.44	.44	.33
PGY3 PD rank-order score (n=59)	.58	.53	.60	.62

To assess predictive ability we conducted two regression analyses, one for PGY3 ITER scores and one for PGY3 PD rank-order scores, as seen in Table 3. ITER scores for PGY1 and PGY2 explained approximately 25% of the variance in PGY3 ITER scores. Adding in the comment ranking scores did not significantly improve prediction. When predicting the PGY3 PD rankings, the PGY1 and PGY2 ITER scores explained much more of the variance, approximately 46%, and this again did not improve significantly with the addition of the comment scores.

Table 3. Results of Regression Analyses to Explore How Several Variables Predict Postgraduate Year Three (PGY3) In-Training Evaluation Report (ITER) Score and PGY3 Program Directors' (PD) Rank-Order Score, University of Toronto Faculty of Medicine, 2009-2011*

	PGY3 ITER score				PGY3 PD rank-order score			
Predictor	*R*	R^2	Delta	*P*-value	*R*	R^2	Delta	*P*-value
ITER1	0.444	0.197	--	--	0.583	0.340	--	--
ITER1 + COM1	0.478	0.228	0.031	ns	0.605	0.366	0.026	ns
ITER2	0.438	0.192	--	--	0.604	0.365	--	--
ITER2 + COM2	0.440	0.193	0.001	ns	0.659	0.434	0.069	†
ITER1	0.444	0.197	--	--	0.583	0.340	--	--
ITER1 + ITER2	0.500	0.250	0.053	†	0.679	0.461	0.121	‡
ITER1/2 +COM1/2	0.521	0.272	0.022	ns	0.706	0.498	0.037	ns
COM1	0.442	0.195	--	--	0.533	0.284	--	--
COM1 +COM2	0.450	0.203	0.008	ns	0.654	0.428	0.144	‡
COM1/2 +ITER1/2	0.521	0.272	0.069	ns	0.706	0.498	0.070	†

* ITER1 = average PGY1 ITER score; COM1 = PGY1 comment score; ITER2 = average PGY2 ITER score; COM2 = PGY2 comment score; ns = not significant.

† Indicates significant increase in R^2 at $P < .05$.

‡ Indicates significant increase in R^2 at $P < .01$.

Discussion

In-training evaluations of residents by faculty are often criticized for poor reliability and validity, and educators are discouraged from using them for summative assessment.[1,12,13] The narrative comments on these forms may hold promise by adding additional information over and above the scores, potentially allowing identification of residents in difficulty that the scores alone do not identify. This study was designed to systematically analyze the written comments on residents' evaluation forms in an attempt to quantify their added value. On post-exercise debriefing faculty attendings reported finding the ranking task easy to complete. Even with faculty blinded to ITER score this process yielded excellent 3-rater reliability and rankings that moderately correlated with performance in PGY3. To our surprise, however, we also found ITER scores to be quite reliable and predictive of PGY3 performance based on PD rank-order—so much so that the comments did not add incremental validity to the predictive model.

Why were our ITERs "better" than what the literature would suggest? Our program is not unique or unusual. It is a large residency dispersed over five major teaching hospitals. Our attending physicians do not receive any sort of special or rigorous training on how to use the ITER forms, which follow a fairly typical format of competency dimensions (based on the CanMEDs roles), a global or overall rating, and a box for free-text comments. Our ITERs, in addition to appearing to consist of only two underlying dimensions, should also suffer from all the pitfalls and biases commonly reported: halo and ceiling effects, leniency, recall bias, etc.[12] Certainly this was true of scores from a single ITER, but across a year's worth of ITER evaluations, the reliability (and predictive validity) of our ITER scores was surprisingly good. That is, while the reliability of a single end of rotation ITER was poor, averaging across 9 rotations yielded reasonable inter-ITER reliability.

A re-examination of the literature suggests that we are not the first to find that ITERs may be quite reliable if the data are aggregated. In 1992, Carline and colleagues[14] reported a study involving 328 IM students over 2 years, whereby multiple evaluations were submitted for each student from various attendings and residents. Their analysis of 3,557 forms (an average of 12 per student) found that 8 forms provided acceptable reliability (using a G-study) when looking at the overall clinical grade assigned. In 1998 Kreiter and colleagues,[15] also using a G-study, found that reliability was acceptable if 3 or more raters evaluated every student on a given rotation. In a surgical residency Littlefield and colleagues[16] analyzed 474 evaluation forms from 35 residents and found that for one rater the reliability would be very low (.13) but for 10 raters it was much higher, at .59. In all of these studies the evaluation forms consisted of multiple items tapping into various elements of performance, all rated on scales (4-7 items) followed by a "global" or overall rating, similar to ours. So the evidence seems to suggest that as long as there are multiple raters, either within or across rotations, there can be acceptable reliability. That begs the question of why educators are so discouraged from using ITERs.[1,12,17]

One reason why ITERs are often discredited relates to the discourse that these ratings are merely "subjective" and therefore somehow substandard to other assessment methods that are purported to be "objective," such as OSCEs.[18] Another issue is that factor analysis studies routinely show that however many items there are on a form they tend to load on two factors, as we found as well, usually along the lines of medical knowledge and interpersonal skills.[19] Indeed, the American Board of Internal Medicine's form was found in one study to load mostly on one single factor.[20] This may serve to undermine people's confidence in what the forms are purported to measure, i.e., multiple competencies,[21] especially as individual practitioners would rarely have access to aggregated scores collected across multiple observers or rotations.

Apart from the apparent reliability of our ITER scores we also found significant reliability in rankings of the ITER comments. When using three raters per resident (and only 15 or16 residents per rater), the reliability of the rank-ordering of comments combined across 63 residents was 0.83 in PGY1 and 0.79 in PGY2. One recent study used similar methodology to explore faculty attendings' ability to rank-order narratives of residents' performance and also found very high inter-rater reliability.[9] One criticism of that methodology was that those narratives were fabricated and covered a much broader range of resident performance than what we usually see in practice. In the current study using real residents' comments over each year of training (and where, incidentally, over 91% of residents receive a 4 or 5 out of 5 on their global rating – see Figure 1) we were still able to show excellent inter-rater reliability, suggesting that experienced faculty attendings have a shared understanding and conceptualization of residents' performance in real-world settings. This echoes recent calls to rely on the wisdom and expertise of assessors.[22, 23]

In the regression models we found that the PGY1 and PGY2 predictors (both numeric and comment scores) were more predictive of the PD rankings than of the PGY3 ITER itself, explaining a significantly higher proportion of the variance. One possible explanation is that the PDs were forced to rank-order their residents, thus separating out many who were likely performing at very similar levels. It is also important to note that although the correlations were high they certainly were not perfect and a lot of the variance in PGY3 performance is left unexplained by PGY1 and PGY2 scores and comments. The rest may be "noise" but it is possible there is more signal yet to be found, such as effects related to rotation type, hospital site, time of year, etc. That comments and scores explained equal amounts of variance can be explained by the high correlation between these two metrics. These unanticipatedly high correlations suggest that raters were documenting their concerns adequately through their commentary and appropriately reflecting their severity through use of the numeric scale when both comments and scores are considered in aggregate. It is also possible that there are more sophisticated ways of analyzing the language on the ITERs that might yield more information and be far less time consuming than having faculty read and rank them. This is being evaluated in current research by our team.

Some limitations of our study include the issue of potential recall bias of the PGY3 PDs, who could also refer to the residents' ITERs if they wished, perhaps confounding the rank-orderings. Also, despite having multiple hospital sites, data were collected from a single institution, which limits generalizability. Further, we acknowledge that although we have focused here on correlations they should be interpreted with caution. As shown in Figure 1, for example, the residents that PDs flagged as problematic were seen to cluster at the bottoms of the rank-orders and tended to have lower than average ITER scores with none higher than 4.3. Using this number as a cut-off or a red flag would

capture many more residents (approximately 40) who are doing well – that is, there would not be sufficient specificity for this to be useful.

Despite these potential concerns, this study represents one of the first we know of to include the analysis of a large cohort of residents followed over a 3-year period using both numeric and narrative data from the ITERs. We also used two different outcome measures (PGY3 ITER scores plus their PD rankings) which were similar, but clearly not the same, given the moderate correlation of 0.59. Although the results bear replication, this systematic analysis of ITER scores and narrative comments does point to the predictive value of ITER scores – despite their bad reputation – as well as offering interesting insights into the potential for the structured use of narrative comments. Although in this study the narrative comments did not offer additional predictive value, our results certainly indicated an impressive amount of "signal" in this data source, and future studies might well seek out mechanisms to automate the quantification of this signal in order to explore opportunities to use this information more effectively.

Acknowledgments: The authors wish to thank Ms. Claire Bouchal for assisting in data preparation and collection, and Mr. Ed Lorens, Research Officer in the Department of Medicine, University of Toronto, for assembling and anonymizing all ITER forms and preparing our primary data set.

Funding/Support: This study was funded by the National Board of Medical Examiners Stemmler Fund for Research in Education.

Other disclosures: None.

Ethical approval: Ethics approval for this study was granted by the Office of Research Ethics, University of Toronto. Protocol #26049.

Previous presentations: This work was presented, in part, as a research poster at the Research in Medical Education Conference at the Association of American Medical College's Annual Meeting. San Francisco, CA. November, 2012.

References

1. Chaudhry SI, Holmboe ES, Beasley BW. The state of evaluation in internal medicine residency. J Gen Intern Med. 2008; 23:1010-1015.
2. Cleland J, Knight LV, Rees CE, Tracey S, Bond CF. Is it me or is it them? Factors that influence the passing of underperforming students. Med Educ. 2008; 42:800-809.
3. Bullock G, Sherbino J, Hodder R, Watling C. Assessment: A White Paper prepared by the Royal College of Physicians and Surgeons of Canada, Future of Medical Education in Canada. Ottawa, Canada: Royal College of Physicians and Surgeons of Canada, 2011.
4. Cohen G, Blumberg P, Ryan N, Sullivan P. Do final grades reflect written qualitative evaluations of student performance? Teach Learn Med. 1993; 5:10-15.
5. Richards SH, Campbell JL, Walshaw E, Dickens A, Greco M. A multi-method analysis of free-text comments from the UK general medical council colleague questionnaires. Med Educ. 2009; 43:757-766.
6. Durning SJ, Hanson J, Gilliland WM,J.M., Waechter D, Pangaro LN. Using qualitative data from a program director's evaluation form as an outcome measurement for medical school. Mil Med. 2010; 175:448-452.
7. Frohna A, Stern DT. The nature of qualitative comments in evaluating professionalism. Med Educ. 2005; 39:763-768.
8. Ginsburg S, Gold W, Cavalcanti R, Kurabi B, McDonald-Blumer H. Competencies "plus": The nature of written comments on internal medicine residents' evaluation forms. Acad Med. 2011; 86:s30-s34.
9. Regehr G, Ginsburg S, Herold J, Hatala R, Eva KW, Oulanova O. Using "standardized narratives" to explore new ways to represent faculty opinions of resident performance. Acad Med. 2012;87:419-27.
10. Frank JR. The CanMEDS 2005 Physician Competency Framework. Better standards. Better physicians. Better Care. Ottawa: The Royal College of Physicians and Surgeons of Canada, 2005.
11. Shrout PE, Fleiss JL. Intraclass correlations - uses in assessing rater reliability. . Psychol Bull. 1979; 86:420-428.
12. Gray JD. Global rating scales in residency education. Acad Med. 1996; 71:S55-S63.
13. Williams RG, Dunnington GL, Klamen DL. Forecasting residents' performance--partly cloudy. Acad Med. 2005; 80:415-422.
14. Carline JD, Paauw DS, Thiede KW, Ramsey PG. Factors affecting the reliability of ratings of students' clinical skills in a medicine clerkship. J Gen Intern Med. 1992; 7:506-510.
15. Kreiter CD, Ferguson K, Lee WC, Brennan RL, Densen P. A generalizability study of a new standardized rating form used to evaluate students' clinical clerkship performances. Acad Med. 1998; 73:1294-1298.

16. Littlefield J, Paukert J, Schoolfield J. Quality assurance data for residents' global performance ratings. Acad Med. 2001; 76:S102-S104.
17. Epstein RM. Assessment in medical education. N Engl J Med. 2007; 356:387-396.
18. Eva KW, Hodges BD. Scylla or Charybdis? Can we navigate between objectification and judgement in assessment? Med Educ. 2012; 46:914-919.
19. Silber CG, Nasca TJ, Paskin DL, Eiger G, Robeson M, Veloski JJ. Do global rating forms enable program directors to assess the ACGME competencies? Acad Med. 2004; 79:549-556.
20. Haber R, Avins A. Do ratings on the American Board of Internal Medicine resident evaluation form detect differences in clinical competence? J Gen Intern Med. 1994; 9:140-145.
21. Lurie SJ, Mooney CJ, Lyness JM. Measurement of the general competencies of the accreditation council for graduate medical education: A systematic review. Acad Med. 2009; 84:301-309.
22. Crossley J, Johnson G, Booth J, Wade W. Good questions, good answers: Construct alignment improves the performance of workplace-based assessment scales. Med Educ. 2011; 45:560-569.
23. Ginsburg S. Respecting the expertise of clinician assessors: Construct alignment is one good answer. Med Educ. 2011; 45:546-548.

CHAPTER 3

Reading Between the Lines: Faculty's Interpretations of Narrative Evaluation Comments

Published as: Ginsburg S, Regehr G, Lingard L, Eva KW. Reading Between the Lines: Faculty's Interpretations of Narrative Evaluation Comments. *Med Educ*. 2015;49(2):296-306.

Abstract

Purpose

Narrative comments are used routinely in many forms of rater-based assessment. Interpretation can be difficult due to idiosyncratic writing styles and disconnects between literal and intended meaning. Our purpose was to explore how faculty interpret and make sense of the narrative comments on residents' In Training Evaluation Reports (ITERs) and to determine the language cues that appear to be influential in generating and justifying their interpretations.

Methods

24 internal medicine (IM) faculty each categorized a sub-group of PGY1 and PGY2 IM residents based solely on ITER comments. They were then interviewed to determine how they made their judgments. Constant comparative techniques from constructivist grounded theory was used to analyze the interviews and develop a framework to understand how ITER language was interpreted.

Results

The overarching theme of "reading between the lines" explained how participants read and interpreted ITER comments. Scanning for "flags" was part of this strategy. Participants also described specific factors that shaped their judgments: consistency of comments, competency domain, specificity, quantity and context (evaluator identity, rotation type and timing). There were several perceived purposes of ITER comments, including feedback to the resident, summative assessment and other more socially complex uses.

Conclusions

Participants made inferences based on what they thought evaluators intended by their comments and seemed to share an understanding of a 'hidden code'. Participants' ability to "read between the lines" explains how comments can be effectively used to categorize and rank-order residents. However, it also suggests a mechanism whereby variable interpretations can arise. Our findings suggest that current assumptions about the purpose, value and effectiveness of ITER comments may be incomplete. Linguistic pragmatics and politeness theories may shed light on why such an implicit code might evolve and be maintained in clinical evaluation.

There are many circumstances in health professions education in which narrative commentary on a trainee's performance plays an influential role. For example, on ward-based in-training evaluation reports (ITERs), which are ubiquitous in postgraduate medical education, comments do not merely communicate information to trainees that could allow them to understand their strengths and weaknesses and improve their performance.[1] These same ITER comments also serve more evaluative purposes such as communicating to the program director information that can support decisions about promotion and remediation. [2] In other contexts, such as the Deans' letters that accompany post-graduate training applications, narrative commentary may be drawn upon to gain a rich understanding of the individual described, such that decisions can be made about who should be admitted to particular training posts or what formal requirements should be put in place regarding continued professional development. These uses of experts' subjective, narrative comments regarding trainee performance – described by Hodges as "post-psychometric" approaches to evaluation[3] – have recently been put forward as "indispensable for trustworthy decision making in summative assessments".[4] However, given that individuals' writing styles are idiosyncratic and given that language in general is often ambiguous, the same written information can be interpreted in multiple ways. Thus, it is difficult to know how "trustworthy" the interpretation of such comments actually is.

The complexity of interpreting narrative comments is well documented in medical education. For example, studies that have looked critically at the language used by clinical supervisors when assessing trainees have found that commonly used positive words can have unexpected meanings. In one study of Deans' letters of applicants applying to an Emergency Medicine residency program, researchers found that the word "good" was only used by 30% of schools; moreover, all schools that did use it reserved the term for the bottom half of the class. [5] The authors concluded that the word "good" was actually a code word for "below average". In another study of application packages to a Radiology residency, the word "excellent" was never used by medical schools as the top category of students and, for more than half the schools, an "excellent" student could be in the bottom half of the class.[6] Despite such vagaries, it has been seen in other contexts that raters can reliably rank-order trainees based on narrative commentary alone.[7] Further, it appears that ITER comments may have value in predicting performance or the need for remediation.[8,9] Put together, such studies suggest that there may be a relatively well-understood 'hidden code' involved in writing and deciphering assessment language.

Theory from the branch of linguistics known as pragmatics can help us understand how features of language beyond literal meaning are used for communication. Well-known examples of non-literal communication, which is common in English, include irony, sarcasm and metaphor. [10] The ability to correctly interpret these non-literal meanings depends heavily on context, including awareness of who is speaking, to whom, with what tone of voice, in what setting, for what purpose, etc. Because ITER comments are a form of linguistic communication in which the reader does not have direct access to many potentially valuable contextual cues, it is important to understand how the language is

being interpreted. Which cues are being picked up by readers and how do they modify the literal meaning, if at all? How is sense made of these types of comments?

In this paper we report findings from interviews with faculty attendings in Internal Medicine (IM), in which we sought to explore faculty's reasoning processes when evaluating ITER comments. Our purpose was to understand how faculty interpret and make sense of the narrative comments written on residents' ITERs and to determine the language cues that appear to be influential in generating or justifying their interpretations. Cracking "the code" of narrative commentary in medical education has the potential to yield considerable insight into assessment practices, their validity, and optimal strategies for faculty development.

Methods

The data collected for the analysis described here were generated during interviews of participants immediately after they had completed a "narrative ranking" task as described in full in Ginsburg et al.[7] That process is described here briefly to provide context for the current analysis. Research ethics board approval was received from the University of Toronto, Office for Research Ethics.

Materials

Each resident in our IM program receives approximately 8-9 ITERs per year, each of which contains 19 items rated on a 5-point scale and a box for free-text comments that states: "Provide a general impression of the trainee's development during this rotation, including general competence, motivation and consultant skills. Please emphasize strengths and areas that require improvement." For the 63 residents in the cohort finishing PGY3 in 2011, comments were present in over 90% of the ITERs and took the form of a few short sentences, point-form statements or entire paragraphs. From these comments, we created two text documents for each resident - one consisting of all ITER comments from the PGY1 year and one consisting of all comments from the PGY2 year. We removed all references to residents' identity including name and gender as well as type of rotation (e.g., name of sub-specialty, hospital site, etc.) For the PGY1 and PGY2 documents separately, the 63 documents were assigned to 12 packages of 15-16 each such that no two packages were alike and each document appeared in three packages. The decision to include 15-16 documents for each rater was based on previous work indicating that this a reasonable number of narratives to be categorized and rank-ordered within a time-frame considered appropriate by participants(8); having three raters per resident document for each PGY set resulted in a required sample size of 24.

Participants and Procedure

To be included in the study, physician participants had to have attended on an inpatient IM service at any of our University's teaching hospitals and had to have had at least two years' experience evaluating residents. This led to a list of approximately 60 eligible faculty from which we recruited 24 attending physicians. The resulting sample

contained 14 men and 10 women, with an average number of years of experience of 9.3 (range 2-33).

In a one-on-one setting, participants were oriented to the four categories describing residents' performance that were developed in a previous study: A=outstanding, excellent, exemplary; B=solid, safe, may need some fine tuning; C=borderline, bare minimum, remediable; D=unsafe, unacceptable, multiple deficits.[7,11] Their first task was to categorize the 15-16 residents in their package by placing as many in each category as they wished. They were then asked to rank-order the residents within each category.

Subsequent to this process each participant was interviewed by a single research assistant who had qualitative research experience in education but who was not involved in any way with our residency program and was thus unknown to participants. One pilot interview was co-conducted with the lead author, but because no changes were made to the protocol afterwards this interview was also included in our dataset. During each semi-structured interview participants were asked about the ranking process, how they decided to put residents into the four categories and to rank-order them, how cut-point decisions were made (i.e., how they decided if a resident was at the bottom of one category or the top of another), and what language in the comments influenced their decisions. They were also asked to provide comments on the ITERs in general. The entire task took approximately 90 minutes per participant with the interview portion lasting between 15 and 30 minutes. Interviews were audio-taped, transcribed and anonymized.

Analysis

The transcripts were analyzed using principles of constructivist grounded theory.[12] As sensitising concepts, we considered that participants may have been influenced by such factors as the strength of adjectives used, the mention of particular competency domains, and the presence of "lukewarm" language which may be interpreted negatively. [13] AA conducted the primary analysis using a line-by-line approach to identify codes that were then grouped into themes. We used a constant comparative approach to coding in an iterative fashion, with each transcript being read numerous times to look for confirming/disconfirming examples in a process that continued until the coding structure appeared stable and sufficient (i.e., until no new codes emerged after multiple reads).[14] The codebook (the coding framework with definitions and examples) was then presented to three other members of the research team along with several un-coded transcripts. Each team member read the transcripts before reviewing the codebook and provided critical feedback on the codes and their interpretation. No substantive changes to the coding were made during this process; rather, feedback was used to further clarify and define existing codes. NVivo 10 was used to organize the data and facilitate coding.[15]

Results

Analysis of the 150 pages of interview transcripts resulted in several themes that provide a framework for understanding how participants came to their rank-ordering and categorization judgments. The overarching theme, which explains *how* participants read and interpreted the ITER comments, we called "Reading between the lines". This strategy will be explored, followed by a description of the specific factors that participants claimed had influenced their rank-ordering of the residents. (Table 1)

Reading between the lines

All participants either directly or indirectly expressed a need to read between the lines when attempting to understand narrative comments:

> They still use a lot of those words, "excellent, outstanding"' but you can read through the lines that [these residents are] not the ones that stand out. (I20).

The word "interpret" or its variations was common in participants' responses: "I think sometimes reading these things does take a little bit of interpretation." (I18). Participants also noted euphemisms, such as: "'Solid' is essentially a euphemism for middle of the road" (I9). Some commented that what appears to be "good" is actually "bad":

> On the face of it, words like "very good", "solid', "at level" can all look good, and solid, and at level, but actually mean "below level, needs work." (I13)

Similarly, " '*Met expectations'*, ironically enough in this kind of evaluation, really means below expectations" (I13), and "that word 'good', to me means someone is trying to be really nice and downplay some deficiencies." (I20) The data abounded with such descriptions of how language should not be taken at face value – that the real meaning was implicit:

> The other thing is that there are some code words, or things that seem like, you know, it says "good" or it says "solid" or it says "strong" – things that you think would be actually good comments, but having read a lot of letters, you realise those are kind of faint praise. (I17)

In sum, these examples demonstrate that language was not taken at face value and that there is an implicit code that was shared, with participants 'translating' words consistently based on their past experience with similar comments.

The specific factors that fed into this code will be explored below. Beforehand, however, it is important to note that the decoding of comments was an active process with participants indicating that they sought particular language cues. They frequently mentioned scanning for "red flags", both positive and negative, to help them find the relevant cues in a sea of comments. For example, one participant said "after a while the comments start to run together a little bit. …you kind of have to look for the important positives or important negatives" (I18). Another explained that:

> [...the] same terms were used over and over again in the comments, such that if there are any different terms, it stood out, whether it was bad or good. (I21)

There were numerous red flag words or phrases consistently identified by participants, suggesting either potential problems (e.g., 'good', 'solid') or superstars ('exemplary', 'chief resident material'). (See Table 2 for further examples).

Specific factors influencing judgments

As participants read between the lines of the comments, several specific factors in addition to the language cues described above appeared to influence their judgments (Table 1).

Consistency

Participants regularly reported being influenced by the consistency of the comments, with every interview containing multiple references to consistencies over time, across different rotations and evaluators, or across domains. In explaining why residents were placed in the A (exemplary) category, one participant noted, "In both cases … they really just had a well-developed knowledge base; multiple evaluators noted that they were above the expected level" (I15). Similarly, supporting their decision regarding a lower-ranked resident, a participant noted that "in every single rotation, there's some hint of extra work that needs to be done." (I13) In these examples, the "multiple evaluators" and "in every single rotation" signaled consistent performance. Consistency of performance across domains was also important to participants, one of whom noted:

> If the comments are always single category comments, I'm less likely to rank them high, whereas if they have multiple competencies [noted to be strong] I'll be more likely to rank them high. (I11)

Although the presence of consistently positive comments across rotations and domains was interpreted favourably, participants' interpretations of inconsistency in comments varied. Inconsistencies were a concern for some participants, suggesting that the resident might be weaker overall. For others, inconsistencies seemed expected: "We all have things we can work on. They had definite deficiencies, but is it bad that they have deficiencies? No, if they're areas that can be improved upon." (I20) Participants expressed discomfort when trying to reconcile inconsistencies, and described divergent strategies for this, including disregarding inconsistent comments as "just one isolated opinion amongst a lot of other superlatives" (I13) or treating "outlier" comments as a signal that needs "to be taken seriously" (I18). Exactly how to reconcile inconsistencies could be a source of tension. For example, one participant struggled to explain why he/she didn't put a resident in category D despite negative comments, finally conceding: "because *someone* thinks they're really good". (I2)

Competency domains

The domain of competency featured in a comment was also influential to participants' interpretation and ranking judgments. Comments about knowledge were specifically viewed as markers of excellence, illustrated in the representative assertion that "you can't be an A without outstanding knowledge" (I14). Conversely, 'conspicuous absences' related to knowledge raised suspicions, particularly if a resident had comments about how hard they worked but none about their knowledge base. Indeed, sometimes comments about the 'implicit competency'[16] of work ethic ("hard working", "great effort", etc) were interpreted as "those nice things you say about everyone" (I21) and thus were thought to be particularly unhelpful. However, knowledge was not always the primary trigger for categorizing: as one participant reported, "the first thing I look for is attitudes and values".

Specificity of comments

More specific and detailed comments were interpreted as a sign that the writer really knew and spent time with the resident; therefore, those comments were seen as more credible and carried more weight:

> What moves somebody from a B to an A is really clear descriptive language as to why they excelled. Not *just* the words 'excellent, exceptional, outstanding', but 'excellent, exceptional, outstanding *because* they did this.' (I14)

By contrast, generic-sounding comments were seen as less credible and were perceived as suspect. Participants felt that they could have been written about anybody and thus did not convey any useful information. Common complaints about such comments were: "I had trouble placing the whole stack of B because they're interchangeable. It's all the same language, it's the mushy middle." (I14), and "It's all generic, right? It's just very basic qualities that I think if you're a nice person, you get those … ." (I21) The dislike of generic language may explain the strategy of scanning for red flags. For some, generic comments led to further reading between the lines, potentially resulting in a negative interpretation:

> People don't like to write negative comments at all so the most negative thing you will see is either nothing written, which is probably telling of itself, or writing comments like "hardworking" and "good" without being very specific. (I3)

Table 1. Definitions of factors that participants considered when interpreting ITER comments, along with representative quotations from interviews transcripts.

Factors influencing judgments	**Definition**	**Representative Quotation**
Consistency	Concordance vs. discordance of ITER comments across rotation, across competencies or over time	"The B category … had areas of excellence, but then inconsistencies. So some rotations, they wouldn't perform as well on as other rotations." (I11)
Competency domains	Specific mention of a particular competency represented in the ITER comments as important to judgment (e.g., knowledge, communications skills, etc)	One "made a number of comments about the person's humanism and their interaction with patients and I thought that was exceptional."(I10)
Specificity	Mention of how specific and concrete vs. generic the ITER comments were	"People weren't specific enough. They thought everyone communicated great and talked to family and all of this. It often didn't allow enough of a division." (I15)
Quantity	How many ITER comments available per resident or how much was written in each	"If the evaluator can't put the time to write any more than one line, then they must think that they were neither really, really good or really, really bad. So it was easy for me to say 'Ok they're safe'." (I21)
Context		
Evaluator identity	Expressed the importance of knowing who wrote particular ITER comments	"Because there's quite a few where it's just 'They're wonderful and super and the best ever' and I don't know if that person's 'wonderful and super and the best ever' is better than someone else's 'wonderful, super, and the best ever'. (I4)
Rotation type	Expressed the importance of knowing from which rotation particular ITER comments arose	With "the subspecialty rotations, it's hard to know what to make of it. From what I know of those rotations, I think you could be bad and look average. It would be possible to not notice those problems, for a variety of reasons. I almost discount that." (I10)
Timing	Stability or changes in ITER comments over the year or within individual rotations; any sense of evolution of performance or skills	"So some of them it was just one rotation that looked a little bit different. And often it was actually the earlier comments and I did make a note of that. So you can see that maybe people in the transition to PGY2 aren't as confident or knowledgeable initially as they were later on and you can see that in the comments." (I18)

Table 2. Examples of red flags, both positive and negative, that participants considered as important signals in their interpretation of ITER comments.

Negative red flags	Positive red flags
Solid, good, reasonable, improved/ improving, continues/continuing, listening or responding to feedback, functioning at the expected/appropriate level	Chief resident material, future colleague, want this person in our program, exemplary, super star, top 1% or 5% of peers, indications that resident is functioning at higher level (e.g., a PGY1 noted to be functioning at PGY3 level, or PGY3 functioning as consultant)

Quantity

The quantity of comments for a given resident was often remarked on, but seemed to be more an indication of comment believability than resident quality, as lengthy comments were seen for both outstanding and problematic residents. Longer comments gave the impression that greater effort went into writing them; therefore, they could be interpreted as an indication of how well the resident was known by the writer or how much effort the writer was willing to expend. Quantity was not simply another marker of specificity, as relatively short comments could be quite specific.

Contextual Factors

Three important contextual factors arose that influenced the interpretation of ITER comments: evaluator identity, rotation type and timing.

Regarding evaluator identity, many participants noted that the style of writing could be markedly different between different attendings – some write more, others less; some use flowery language, others are more terse; some use superlative adjectives, others do not. Since the evaluator was not known, and it was unlikely that the same person wrote more than one comment for a given resident, participants found this a frustrating aspect of the research task:

> Even though I know that I'm *aware* that evaluators vary in how they evaluate residents, there is no ability to control for that because you don't know who the people are, how they evaluate other residents. (I19)

Different approaches to writing comments were perceived as potentially being detrimental to residents:

> There's so much variability in the way that people write these comments … some people *always* write in areas for improvement in *every* narrative comment they do and everybody has areas in which they can improve. But if you actually put in an area for improvement it basically makes that resident look worse, especially if 90% of the evaluators are not putting in any areas for improvement. (I19)

This made it difficult for participants to know how much weight to place on areas for improvement and also explained why the identity of the writer was seen as an important interpretive factor.

Many participants also felt that knowing the rotation type was essential to their interpretation. A general internal medicine (GIM) rotation's comments carried more weight than comments from a subspecialty, especially those that have shorter attending blocks: on a GIM rotation "they have a lot of exposure to the residents… so they have a broad base by which to judge them". (I11)

The third contextual factor was timing. For example, many participants thought it was important to note the time of year from which certain comments were derived:

> [It's] a little bit challenging because not all the evaluations are consistent and some of them are *really* good, some of them are not as good, so I tried to figure out as to whether the evaluations were done at the beginning when the evaluation was not that great, or maybe the residents did get better over the 12 month period. (I24)

Lack of improvement over time could suggest a lack of insight on the resident's part. By contrast, participants recurrently commented on the use of verbs indicating change (e.g., improving, developing, continues, evolving) as implying a negative characteristic of resident performance. As I16 said, "'demonstrated tremendous growth' [is] a positive because it shows that they are able to respond to feedback … but at the same time, [this] suggests that their performance wasn't very good."

General comments about the ITERs

To gain further understanding into sources of idiosyncrasy in comment writing and interpretation in this context, participants were asked to discuss their opinions regarding the general purpose and use of ITERs. Many considered the ITER as a means for providing formative feedback so that residents could continue to improve. The word "feedback" arose repeatedly when discussing the purpose of the ITER. Others noted that in practice they provide much more constructive feedback during the rotation or in a discussion setting but don't necessarily document everything on the form, viewing the ITER as purely summative, a "final judgment" of a resident's performance. For some, the ITER was seen as a way to communicate "red flags", knowing that the program director would see them; as one said, they send the message that there's something "that needs to be looked into". (I10) Some felt the purpose of comments on the ITERs was to enable good reference letters to be written as residents apply for fellowships or chief residency positions.

Discussion

The purpose of this analysis was to understand how faculty are able to make sense of narrative comments on resident ITERs and to determine the language cues that shape these interpretations. Previous work has reported that, based on comments alone, participants could assign scores to residents' performance with impressive reliability.[7] We found with the current work that participants shared a strategy of reading between the lines, scanning for flags and making inferences based on what they thought evaluators were trying to convey through the language they used. These strategies influenced their interpretation of the comments with several factors (consistency, competency domain, specificity, quantity and context) being described as important elements of a hidden code.

Despite the apparent existence of shared decoding strategies, the use of coded language was not without difficulty. Our participants claimed to struggle with interpreting vague and generic comments, often focused on residents' disposition, echoing a study by Lye et al which found that the single most common phrase in pediatric clerkship evaluations was "pleasant/pleasure to work with", a finding the authors considered alarming for its irrelevance for success as a medical student.[17] In that study, comments related to specific clinical skills were found only 31% of the time. Similarly, Ginsburg et al, in a content analysis of written comments on IM residents' ITERs, found that comments about a resident's "attitude or disposition" were common, along with other commentary not linked directly to competencies.[16] The problem of writing vague, dispositional comments that are subject to (mis)interpretation is not unique to medicine and can be found elsewhere in higher education. [18–20]

Given their prevalence, it is intriguing to speculate about why assessors write vague and generic dispositional comments when specificity is clearly valued. Faculty development often assumes individuals do not know better or don't take the time to write carefully and, therefore, efforts are aimed at teaching faculty to write "better" comments by being more specific. However, the fact that vague and generic comments were criticized by individuals who were drawn from the very same pool of individuals who wrote them suggests that something is lost in translation (i.e., that the comments might seem more readily interpretable when recorded than they are when read).

Unless there is a shared understanding of (and commitment to) the purpose of an assessment it is difficult to know what sorts of comments most effectively convey the meaning intended by the author. As noted by others [1] and as seen in our data, however, it is likely that the ITER is serving multiple purposes simultaneously, some of which may involve considerable social complexity. One potential social purpose may be a desire to attend to residents' 'face' (i.e., the positive image a person has for him/herself). According to theories of politeness[21], by emphasizing positive skills that are perceived to be of great value to the team – such as being hardworking, pleasant to work with and possessing "those other basic qualities that, if you're a good person, you get" - faculty may be allowing residents to "save face", or to maintain or enhance their positive self-image. It is possible that faculty are able to do this because they

believe readers share the code for interpreting their comments accurately, and thus they can attend to residents' face while still sending their intended message.

A second politeness concept that may be relevant here is known as 'conventional indirectness': the use of phrases that, by virtue of convention, "have contextually unambiguous meanings which are different from their literal meanings"[21]. This can explain why words and phrases such as "good", "solid" and "meets expectations" are understood as intending to convey performance that is borderline or below-average without attendings having to use those undesirable terms. [5,6] Although these meanings seem clear to physician readers, it is important to note that residents' interpretations of these terms are unknown. If residents take the terms at face value, they may not appreciate the degree to which their performance could be improved. If they do not take the terms at face value, the cost of residents understanding the code could be that they lose the 'face' that the faculty seek to help them preserve.

In either case, the data collected in this study clearly indicate that, while it is generally useful, the code is not universal and is difficult to decipher without a full understanding of the author's context. As an illustration of this, consider the wording on our ITERs' comment box instructing attendings to discuss the residents' strengths and weaknesses (presented in the methods). It echoes guidelines for writing good ITER comments that stress the importance of documenting areas of strength and weakness along with the resident's response to feedback.[22] While balanced, constructive written feedback may arguably lead to improved resident performance, it might also work to their detriment. Our participants picked up on language cues indicating areas for improvement (or previously enacted change) as reflective of a weaker performance overall. This of course raises issues for resident education. If there is a hidden code but it is imperfectly understood and applied, a resident could look bad if they don't improve but could look equally bad if they do. Or, more to the point, they could look bad if their improvement is documented. This highlights the problem of misalignment between the intended purpose of the ITER and its actual (or perceived) use. For any assessment instrument, such misalignment can increase the risk of "arbitrary judgment",[23] and thus it is of critical importance to understand *how* the instrument is actually being used and interpreted. Indeed, our participants expressed concern that without knowing who the supervisor was (and whether, for example, s/he documents areas for improvement for all trainees), they were not certain how to interpret these "balanced" comments.

This leads to consideration of an additional concept from linguistics that may help explain the frustration attendings expressed over not having full knowledge of the context from which the comments arose. Linguistic pragmatists have labelled the idea that contextual information is necessary to understand the meaning of certain words and phrases as deixis. One type of deixis refers to knowledge of person, place or time as essential for understanding a narrative.[24] Our participants routinely expressed a desire for more information along these lines and felt that, in their absence, they couldn't properly assess the comments. However, this may speak more to their confi-

dence in rank-ordering the residents than to their actual abilities to do so (i.e., these "deictic markers" may be more of a perceived necessity rather than an actual need).

In sum, the multiple apparent purposes expected of ITER comments, the idiosyncratic faculty writing styles, and the absence of what is felt to be key information in many ITER comments make it surprising that participants (as demonstrated in previous work) were able to reliably rank-order residents based on comments alone.[7] Their strategy of reading between the lines and decoding the written comments appears to have been remarkably consistent across preceptors. Nevertheless, more practical research is required to explore this further, including the important questions of (a) whether residents also know the code for reading between the lines effectively, (b) the extent to which the code may be locally specific (e.g., to a particular discipline or residency program) vs universal , (c) whether the message consistently received was in fact the message that was intended and (d) how much written commentary (i.e., how many rotations' worth of data) is required to form stable impressions of residents using this decoding mechanism. Further exploration and application of theories from linguistics (including deixis and politeness) may be particularly fruitful.

Limitations

Our ability to tease apart some of these interpretations is limited by design constraints (e.g., single institution and program). Further refinement of our understanding will benefit from more diverse sampling to assess the degree of institutional specificity of the 'code' and whether residents would read between the lines and interpret the comments in the same way as faculty. If they do, then perhaps the lack of specificity in faculty commentary is not as important an issue as is currently suggested. Finally, it is conceivable that a participant recognized his or her own narratives in the documents provided, thus creating the potential that some participants told us of cues they try to embed rather than cues they can meaningfully observe in others' writing (although no comments of this nature arose in the interviews).

Conclusions

This study is the first, to our knowledge, to explore *how* faculty interpret written comments about residents, and provides an important step towards the goal of using "rich, narrative evaluations of performance" for "trustworthy decision making" in assessment.[4] Participants' ability to "read between the lines" explains how they made sense of written comments and how they were able to effectively categorize residents. However, this strategy also suggests a mechanism whereby variable interpretations can easily arise, particularly when contextual information is missing and inferred. Current assumptions about the effectiveness of ITER comments may be incomplete; thus it is important to conduct additional research. Linguistic pragmatics, including theories of politeness and the concept of deixis may shed light on why such an implicit code exists in written assessment language and how it accomplishes its social functions. Once a more sophisticated understanding of the powerful, yet complex practice of comment

writing and interpretation is established, then effective faculty development initiatives can be developed.

Acknowledgments: The authors wish to thank Professor Cees van der Vleuten for providing constructive feedback on this paper. We also thank Claire Bouchal for conducting the interviews and Lisa St. Amant for assistance with analysis.

Funding/Support: This study was funded by the National Board of Medical Examiners Stemmler Fund for Research in Education.

Ethical approval: This study was approved by the Office of Research Ethics, Faculty of Medicine, University of Toronto.

References

1. Watling CJ, Kenyon CF, Schulz V, Goldszmidt MA, Zibrowski E, Lingard L. An Exploration of Faculty Perspectives on the In-Training Evaluation of Residents. *Acad Med*. 2010;85(7):1157-1162.
2. Cohen G, Blumberg P, Ryan N, Sullivan P. Do final grades reflect written qualitative evaluations of student performance? *Teach Learn Med*. 1993;5(1):10-15.
3. Hodges BD. Assessment in the post-psychometric era: Learning to love the subjective and collective. *Med Teach*. 2013;35(7):564-568.
4. Govaerts MJB, van der Vleuten CPM. Validity in work-based assessment: expanding our horizons. *Med Educ*. 2013;47(12):1164-1174.
5. Kiefer CS, Colletti JE, Bellolio MF, et al. The "Good" Dean's Letter. *Acad Med*. 2010;85(11):1705-1708.
6. Naidich JB, Lee JY, Hansen EC, Smith LG. The Meaning of Excellence. *Acad Radiol*. 2007;14(9):1121-1126.
7. Ginsburg S, Eva KW, Regehr G. Do in-training evaluation reports deserve their bad reputations? A study of the reliability and predictive ability of ITER scores and narrative comments. *Acad Med*. 2013;88(10):1539-1544.
8. Schwind CJ, Williams RG, Boehler ML, Dunnington GL. Do individual attendings' post-rotation performance ratings detect residents' clinical performance deficiencies? *Acad Med*. 2004;79(5):453-457.
9. Guerrasio J, Cumbler E, Trosterman A, Wald H, Brandenburg S, Aagaard EM. Determining need for remediation through postrotation evaluations. *J Grad Med Educ*. 2012;4(1):47-51.
10. Akmajian A, Demers R, Farmer A, Harnish R. Ch. 9 Pragmatics. In: Akmajian A, Demers R, Farmer A, Harnish R, eds. *Linguistics. An Introduction to Language and Communication*. Sixth. Cambridge, MA: MIT Press; 2010:363-418.
11. Regehr G, Ginsburg S, Herold J, Hatala R, Eva KW, Oulanova O. Using "Standardized Narratives" to Explore New Ways to Represent Faculty Opinions of Resident Performance. *Acad Med*. 2012;87(4):419-427.
12. Charmaz K. Coding in grounded theory practice. In: *Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis*. Vol 1st. London, UK: Sage Publications; 2009:42-71.
13. Frohna A, Stern DT. The nature of qualitative comments in evaluating professionalism. *Med Educ*. 2005;39(8):763-768.
14. Dey I. Chapter 6. Concluding. In: *Grounding Grounded Theory: Guidelines for Qualitative Inquiry*. San Diego: Academic Press; 1999:116-150.
15. Ltd. QSRIP. NVivo qualitative data analysis program. 2013;10.
16. Ginsburg S, Gold W, Cavalcanti RB, Kurabi B, McDonald-Blumer H. Competencies "Plus": The nature of written comments on internal medicine residents' evaluation forms. *Acad Med*. 2011;86(10 Suppl):s30-s34.
17. Lye PS, Biernat KA, Bragg DS, Simpson DE. A Pleasure to Work With: An Analysis of Written Comments on Student Evaluations. *Ambul Pediatr*. 2001;1(3):128-131.
18. Giles TM, Gilbert S, McNeill L. Nursing students' perceptions regarding the amount and type of written feedback required to enhance their learning. *J Nurs*

Educ. 2014;53(1):23-30.
19. Agius NM, Wilkinson A. Students' and teachers' views of written feedback at undergraduate level: a literature review. *Nurse Educ Today*. 2014;34(4):552-559.
20. Hyland K. Student perceptions of hidden messages in teacher written feedback. *Stud Educ Eval*. 2013;39(3):180-187.
21. Brown P, Levinson SC. *Politeness : Some Universals in Language Usage*. (Levinson SC, ed.). New York: Cambridge University Press; 1987.
22. Dudek NL, Marks MB, Wood TJ, Lee AC. Assessing the quality of supervisors' completed clinical evaluation reports. *Med Educ*. 2008;42(8):816-822.
23. Berendonk C, Stalmeijer RE, Schuwirth LW. Expertise in performance assessment: Assessors' perspectives. *Adv Heal Sci Educ*. 2013;18(4):559-571.
24. Levinson SC. Chapter 5. Deixis. In: Horn LR, Ward G, eds. *The Handbook of Pragmatics*. Vol 2nd. Oxford, UK: Blackwell Publishing; 2006:97-122.

CHAPTER 4

The Hidden Value of Narrative Assessment Comments: A quantitative reliability analysis of qualitative data

Under peer review

Abstract

Purpose

In-training evaluation reports (ITERs) are the predominant methods of assessment in internal medicine (IM) residency. Most of the focus is on their numeric scores, yet the written comments can provide a rich source of data. Our main goal was to determine the reliability of using variable amounts of commentary to rank-order residents.

Methods

ITER comments were collected from two cohorts of PGY1s in IM at the University of Toronto (n=48 and 46). Packages of comments from 15-16 PGYs from both one full year and from the first 3 months were collated and were rank-ordered by 24 faculty internists across Canada, external to our institution. Generalizability analyses were conducted using G_string.

Results

For the full year of comments reliability coefficients averaged across 4 raters were G=0.85 and G=0.91 for the first and second cohort, respectively. For a single rater values were G=0.60 and G=0.73. When only the first 3 months of comments were used the reliabilities remained high at G=0.66 and G=0.60 for a single rater. Comparable G coefficients for the numeric scores were between 0.08 and 0.19. A decision study for 2 raters and 3 months of comment data showed reliability coefficients of G=0.80 and G=0.75.

Conclusions

Using narrative comments alone to discriminate between residents can be extremely reliable even after only a few reports are collected. This can provide a way to assist residents in difficulty early on, with the advantage (as compared to numeric scores) of offering insight into what needs to be improved.

The assessment of competence in medical education is undergoing a significant transformation. Steps are being taken to prioritize outcome (or competency based) models rather than strictly time-based ones, which will necessitate a major shift in how we assess our trainees.[1] In order to ensure that residents meet predetermined milestones[2] it will become necessary to collect much more information on each trainee to support valid judgment and decisions. [3,4] To this end, increasing value is being placed on qualitative and subjective data[5] and on the need to aggregate data from multiple, low-stakes sources.[6] The format and variety of evaluations is expanding in step with these changes but we still rely heavily on end-of-rotation assessment forms (herein called in-training evaluation reports, or ITERs) that contain numeric and narrative data. Most of the literature (and use) of forms of this type is based entirely on the numeric data yet there may be great value in the narratives.[7–9] Researchers and educators have called for the medical education community to "expand our horizons" of assessment and go beyond numeric ratings to incorporate qualitative and other forms of data.[10]

Some research has been conducted on the utility and feasibility of using assessment comments to evaluate learners or practitioners yielding mixed results. For example, a couple of studies have found that comments are usually concordant with scores assigned, suggesting that reading thousands of comments (e.g., for MD revalidation[11] or residency[12]) may not be worth the trouble. On the other hand, areas of non-concordance can illustrate weaknesses not otherwise picked up by the scores,[7,13] thereby helping to overcome the well-described phenomenon of "failure to fail"[14,15] and providing learners with more guidance regarding how to improve.[16,17] Determining how to balance these competing issues of gaining additional information and maintaining feasibility is an important challenge for the health professions to address, not only in formal training environments but across the continuum of training and practice.

Recent work has found that comments can provide a highly reliable way to distinguish between residents in Internal Medicine even in the absence of numeric scores. Those analyses, however, used data from an entire year's worth of ITER comments for each resident and the comments were assessed by faculty who worked in the same training program as the residents.[7,18] The high reliability observed, therefore, may be related to the volume of comments amassed (given that reliability can generally be expected to increase with the amount of data available in any assessment process) and to the faculty's awareness of the culture within the particular training program studied. Waiting an entire year for evaluations to accumulate would severely limit the usefulness of narrative assessment for early intervention. Being dependent on raters who are intimately familiar with the context would similarly limit the capacity to use such assessments for a variety of purposes.

The goal of this study was to further test the validity of using narrative data in assessment by determining the reliability when using variable amounts of commentary as a means to rank-order residents. We used two cohorts of PGY1 trainees in an internal medicine (IM) program to determine the comparability of reliabilities achieved if the narrative data consist of only comments received early in the year relative to utilizing a full year's sample. In addition, we recruited faculty who did not work in the same program as the residents, but rather, were drawn from different institutions and universities at a national level.

Methods

Setting

After receiving Research Ethics Board approval, ITERs were collected from two cohorts of IM residents in their PGY1 year at one large IM program (total number of residents in each year was 55-56). Each resident receives one ITER at the end of each one-month clinical rotation, over 93% of which contain written comments. Our goal was to include 48 residents from each of two cohorts who had received 8 or more ITERs containing comments over the course of one year, 3 of which had to come from the first 4 months of training. We chose 8 based on studies showing acceptable reliability of ITER scores aggregated across this number of ITERs.[7,19,20] However, in the older cohort there were fewer residents who had ITERs with comments (both overall and from the first 4 months) so we included 3 who had fewer than 8 (two had 7, one had 6).

Numerically, the ITERs contain 18 items, each rated on a scale from 1-5 followed by an overall rating and a free-text box in which to enter comments. After anonymizing the comments and removing the numeric data we created a document for each resident that contained the entire years' worth of comments. See Box 1 for an example. We then created a parallel set of documents containing only the comments from the residents' first three comment-containing ITERs that were gathered in the first 4 months of training. To enable participants to read and rank order residents, twenty-four sets of documents were created for each cohort. To avoid potentially confounding the data by inadvertently grouping higher or lower performing residents together each set contained comments on 15-16 residents and each resident appeared in 3-4 different sets with no two sets being identical.

First three months of comments, Resident # 1234

XXXXXXX is a very pleasant, conscientious, hard-working and reliable housestaff whose knowledge in general internal medicine is excellent. XXXXXXX clinical assessment was thorough and complete. For example, by taking a careful history and followed up with appropriate testing XXXXXXX picked up a case of non-STEMI that was missed by the referring service. During this rotation, XXXXXXX consultative skills have matured nicely and with further experience I expect that XXXXXXX will do very well. Finally, XXXXXXX related well to the healthcare team and XXXXXXX patients. Overall, an excellent performance in a very busy and demanding service.

Dr. XXXXXXX was a pleasure to have on the nephrology service. XXXXXXX did well in the diagnosis and management of both acute and chronic kidney disease. XXXXXXX contributed well to rounds and teaching sessions.

-Great work ethic
-Seemed to end up with many admissions whenever on call and therefore had very large pt load-handled this very well
-complete notes
-thoughtful approach to pt issues
-well done

Box 1. Examples of 3 rotations' ITER comments for one PGY1 resident

Participants

We recruited 24 IM faculty from institutions across Canada by accessing publicly available directories on academic department of medicine websites and emailing study invitations. We also displayed recruitment notices at national medical education conferences and meetings, sent open invitations via twitter, and encouraged word-of-mouth referrals. Potential participants had to have at least two years' experience teaching and evaluating residents on IM Clinical Teaching Units (CTUs). There were no specific exclusion criteria. Study design is shown in Figure 1. Each consenting faculty member was sent a package containing two sets of data: one set contained the entire year worth of comments for 15-16 residents in one of the two cohorts; the other set contained the first 3 months of comment data from 15-16 different residents from the other cohort.

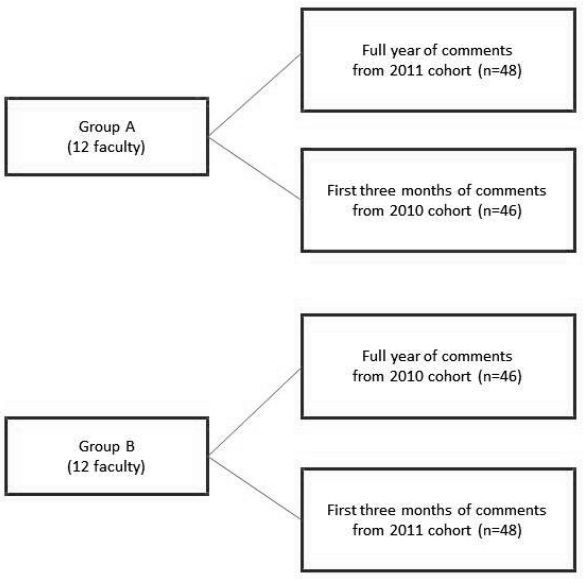

Figure 1. Representation of study design

Protocol

A trained research assistant (RA) conducted a face-to-face meeting over Skype with each participant. Beginning with the set of full-year comments, participants were asked to read all 15-16 documents and sort them into categories derived during prior research (A = outstanding, excellent, exemplary; B = solid, safe, may need some fine tuning; C = borderline, bare minimum, remediable; D = unsafe, unacceptable, multiple deficits).[21] Afterwards they were asked to rank-order the residents within each category, resulting in a final ranking of 1-15/16. After this task they were interviewed by the RA to explore their decision-making process. Following the interview they repeated the task using the second set of documents, which contained a different set of residents' comments from the first 4 months of the year. Time required for the full-year and part-year tasks was approximately 45 and 15-20 minutes, respectively. In total, each resident's comments were expected to be rated by 4 faculty within each condition.

Analysis

To analyze the ITER numeric data we averaged the ratings from the 18 items and the global rating to create a single score per ITER. To analyze the effectiveness of generating judgments based on ITER comments, rank-order data from all 24 faculty partici-

pants was entered into excel and verified for accuracy. The reliability of both ratings and resident rankings was assessed using generalizability theory, with rater nested within resident. G_string was used as it enables analysis when study design is unbalanced (residents had an average of 10 ITERs each with a range of 8 to 11). For the ITER comments the design was also unbalanced as some PGY1s had 3 raters while most had 4 each (this occurred because one resident's package was inadvertently miscoded and, therefore, was deleted from the final results). While the generalizability theory formulation is the same in each instance, the reliabilities for the ratings and the rankings should not be directly compared because the rating analysis is based on a series of ITERs completed by individuals whereas the ranking analysis examines the reliability of judgments provided by faculty who have reviewed the narrative comments completed by multiple individuals. The reliabilities of the ratings are included for the sake of completeness whereas the reliability of the rankings provides the answer to the research questions posed. Correlations between 3-rotation and full-year data were calculated using SPSS (version 23).

Results

Our 24 participants were affiliated with 7 universities: 5 from the University of Calgary; 4 each from the University of British Columbia, University of Ottawa and Western University; 3 each from University of Alberta and McGill University; and 1 from McMaster University.

ITER numeric data:
The 2011 cohort comprised 48 PGY1 residents each of whom received at least 8 ITERs over the year (mean of 10.0) while the 2010 cohort comprised 46 residents (mean of 10.0 ITERs). The reliability coefficients examining the extent to which residents in 2011 and 2010, respectively, could be consistently differentiated based on a single ITER score were G=0.11 and 0.13 for the full year worth of ITER ratings; G=0.08 and 0.19 for the set of the first 3 assessments. (See Table 1). When ratings were aggregated across the full year worth of data the reliabilities of the resulting scores increased to G=0.56 and 0.59 when the full year worth of ratings was used or to G=0.46 and 0.70 when the first 3 rotations were used for the 2011 and 2010 cohorts, respectively.

Pearson correlations between ITER scores averaged across the 3-rotations and ITER scores averaged over the full year were calculated for each cohort and were r=0.78 and 0.66, respectively, both significant with $p<0.01$. While these correlations are artificially inflated due to the full year average including the first 3 rotation scores, this was done for the sake of comparability given that the rankings (below) based on a full

year's set of comments were not generated without consideration of the comments received on the first three ITERs.

Table 1. Reliability of ITER numeric scores for PGY1 residents from two cohorts, based on data gathered over one full-year and from the first three rotations.

ITER scores	2011, full year		2011, first 3 rotations		2010, full year		2010, first 3 rotations	
Source of Variance	R	I:R	R	I:R	R	I:R	R	I:R
Estimated Variance Component	0.035	0.276	0.018	0.214	0.036	0.247	0.038	0.161
Percentage of Total Variance	11.1	88.9	7.8	92.2	12.8	87.2	19.1	80.9
Reliability for a single rater	0.11		0.08		0.13		0.19	
Reliability based on average of ten ratings	0.56		0.46		0.59		0.70	
Note: R = Resident; I:R = ITER nested within Resident								

ITER comments:
The 48 residents from the 2011 cohort were rank-ordered by an average of 3.97 faculty and the 46 residents from the 2010 cohort were rank ordered by an average of 3.94 faculty. For the full year of comments reliability coefficients averaged across all raters were G=0.85 and G=0.91 for the first and second cohort, respectively. The reliability coefficients examining the extent to which residents could be consistently differentiated based on the rank ordering provided by a single faculty member were G=0.60 and 0.73 for the full year worth of ITER comments; G=0.66 and 0.60 for the set of the first 3 assessments. When rankings were aggregated across multiple faculty rankers the reliabilities increased as illustrated in table 2.

Pearson correlations between 3-rotation and full-year rankings were calculated for each cohort and were 0.69 and 0.66 respectively, both significant with $p<0.01$.

Table 2. Reliability of ranking two cohorts of PGY1 residents based on comments alone from one full year and from the first three rotations

Rankings from Comments	2011, full year		2011, first 3 rotations		2010, full year		2010, first 3 rotations	
Source of Variance	R	F:R	R	F:R	R	F:R	R	F:R
Estimated Variance Component	0.057	0.039	0.064	0.032	0.071	0.026	0.057	0.039
Percentage of Total Variance	59.9	41.0	66.4	33.6	72.8	27.2	59.6	40.4
Reliability for a single ranker	0.60		0.66		0.73		0.60	
Reliability based on average of two rankings	0.74		0.80		0.84		0.75	
Reliability based on average of three rankings	0.81		0.86		0.89		0.82	
Reliability based on average of four rankings	0.85		0.89		0.91		0.85	
Notes: R = Resident; F:R = Faculty ranker nested within Resident. The reliabilities should not be compared across table due to the rankings being based on comparisons of 3 or 10 sets of comments whereas the ratings are based on only a single ITER completed by a single faculty member								

Discussion

Our findings reveal that using narrative comments alone as a means of assessing residents can be extremely reliable. This high reliability was maintained even when we considered only the first three ITERs of the year (see table 2). Indeed, the G coefficients based on the first three months of comments are similar to those found after 10 numeric ratings.[7,22] Further, residents' rankings from the first 3 ITERs were highly correlated with their rankings based on the full year of data (although it must be kept in mind that the full year ranking included the ITER comments collected on the first 3 ITERs, so these measures are not truly independent). Our Decision study showed that a reliability of 0.75-0.8, which is in the range of acceptability for even high-stakes assessments,[23] can be achieved with only two faculty members ranking residents based on three rotations worth of comments. This suggests that a simple intervention – having two faculty read residents' evaluation comments early in the year – can be a very

fruitful enterprise and may pick up residents requiring additional educational supports at an early time point.[17]

Unlike previous work, a strength of this study is that the faculty participants were external to our training program and were not trained in assessing ITER comments although they were experienced in IM assessment. Previous research found that faculty belonging to the same program as the residents whose ITERs were being assessed were adept at "reading between the lines" to decode assessment comments that could often appear to be vague and lacking in specificity.[18,24] The fact that external, untrained faculty could read between the lines just as readily implies that there is a degree of universality to how internal medicine faculty write and understand narratives about their residents. This further suggests that there is a shared understanding on the part of faculty of what performance should look like for PGY1s in internal medicine, at least within a single country. This knowledge can help in the attempt to set expectations and standards for PGY1s in evolving competency-based curricula.[1,25]

Our findings have broad relevance to other assessments that collect words as data, such as "field notes" in family medicine[26] or evaluation of teacher competence.[27] Further, they might help to facilitate the educational advantages of assessment processes that are strived for during the continuing professional development stage of practice, a context in which scores are often not helpful due to the narrow range and positive skew that is commonly reported. Before concluding in these regards our results would require replication in different contexts, but the reality that our comments were easily collected, fairly brief and involved no special training on the part of the attendings makes it easy to envision numerous potential applications.

Several limitations should be kept in mind when interpreting our findings. The replicability of our work in other programs may be limited as all of our assessment comments came from a single, albeit large IM program that might have a specific culture of assessment regarding the extent and nature of comments and because our participants were required to have two years' worth of experience with ITERs. Since we used an open recruitment strategy we cannot state a response rate to our call for participation, nor can we make claims about the representativeness of our sample compared to all academic internists. However, the geographic representation (including faculty from programs of all sizes at seven universities) speaks to the generalizability of our findings.

Conclusions

The incorporation of narrative comments as a routine part of assessment in medical education is overdue.[28] Our study adds to the growing validity evidence for the utility of narratives[29] by demonstrating that they can be reliably used as a way to discriminate between residents after a small number of reports are collected. This is particularly useful knowledge if one hopes to intervene quickly to assist residents in difficulty as comments offer far more in the way of what the resident needs to do in order to improve relative to ratings.[16,30] From a practical point of view, most Internal Medicine programs can easily implement a system in which the first three months of ITER comments are assessed by one or two attendings for purposes of flagging residents who may be struggling and ensuring they are guided towards improvement. From a program perspective, narrative comments can provide insight into common areas of weakness that can then be addressed at a curricular level.

Importantly, these findings should help to dispel the common opinion that ITERs are "useless" for assessment in Internal Medicine, which might further reinforce the importance of writing rich and meaningful comments.

Acknowledgements: The authors are grateful to Ms. Lisa St. Amant for conducting and transcribing the interviews.

Funding: This study was funded by the National Board of Medical Examiners' Edward J. Stemmler Fund for Research in Education.

Ethical Approval: This study was approved the Research Ethics Board of the University of Toronto.

References

1. Royal College :: Competence by Design (CBD). http://www.royalcollege.ca/portal/page/portal/rc/resources/cbme.
2. Iobst WF, Sherbino J, Cate O Ten, et al. Competency-based medical education in postgraduate medical education. *Med Teach*. 2010;32(8):651-656.
3. Caverzagie KJ, Iobst WF, Aagaard EM, et al. The internal medicine reporting milestones and the next accreditation system. *Ann Intern Med*. 2013;158(7):557-559.
4. Williams RG, Dunnington GL, Mellinger JD, Klamen DL. Placing Constraints on the Use of the ACGME Milestones: A Commentary on the Limitations of Global Performance Ratings. *Acad Med*. 2015;90(4):404-407.
5. Hodges BD. Assessment in the post-psychometric era: Learning to love the subjective and collective. *Med Teach*. 2013;35(7):564-568.
6. Schuwirth LW, van der Vleuten CPM. Programmatic assessment: From assessment of learning to assessment for learning. *Med Teach*. 2011;33(6):478-485.
7. Ginsburg S, Eva KW, Regehr G. Do in-training evaluation reports deserve their bad reputations? A study of the reliability and predictive ability of ITER scores and narrative comments. *Acad Med*. 2013;88(10):1539-1544.
8. Guerrasio J, Cumbler E, Trosterman A, Wald H, Brandenburg S, Aagaard EM. Determining need for remediation through postrotation evaluations. *J Grad Med Educ*. 2012;4(1):47-51.
9. Overeem K, Lombarts MJMH, Arah OA, Klazinga NS, Grol RPTM, Wollersheim HC. Three methods of multi-source feedback compared: A plea for narrative comments and coworkers' perspectives. *Med Teach*. 2010;32(2):141-147.
10. Govaerts MJB, van der Vleuten CPM. Validity in work-based assessment: expanding our horizons. *Med Educ*. 2013;47(12):1164-1174.
11. Richards SH, Campbell JL, Walshaw E, Dickens A, Greco M. A multi-method analysis of free-text comments from the UK General Medical Council Colleague Questionnaires. *Med Educ*. 2009;43(8):757-766.
12. Durning SJ, Hanson J, Gilliland W, McManigle JM, Waechter D, Pangaro LN. Using Qualitative Data From a Program Director's Evaluation Form as an Outcome Measurement for Medical School. *Mil Med*. 2010;175:448-452.
13. Frohna A, Stern DT. The nature of qualitative comments in evaluating professionalism. *Med Educ*. 2005;39(8):763-768.
14. Cleland J, Knight L V, Rees CE, Tracey S, Bond CF. Is it me or is it them? Factors that influence the passing of underperforming students. *Med Educ*. 2008;42(8):800-809.
15. Dudek NL, Marks MB, Regehr G. Failure to fail: the perspectives of clinical supervisors. *Acad Med*. 2005;80(10 Suppl):S84-S87.
16. Watling CJ, Kenyon CF, Zibrowski EM, et al. Rules of engagement: residents' perceptions of the in-training evaluation process. *Acad Med*. 2008;83(10 Suppl):S97-S100.
17. Cohen G, Blumberg P, Ryan N, Sullivan P. Do final grades reflect written qualitative evaluations of student performance? *Teach Learn Med*. 1993;5(1):10-15.

18. Ginsburg S, Regehr G, Lingard L, Eva KW. Reading Between the Lines: Faculty's Interpretations of Narrative Evaluation Comments. *Med Educ*. 2015;49(2):296-306.
19. Carline JD, Paauw DS, Thiede KW, Ramsey PG. Factors affecting the reliability of ratings of students' clinical skills in a medicine clerkship. *J Gen Intern Med*. 1992;7(5):506-510.
20. Kreiter CD, Ferguson K, Lee WC, Brennan RL, Densen P. A generalizability study of a new standardized rating form used to evaluate students' clinical clerkship performances. *Acad Med*. 1998;73(12):1294-1298.
21. Regehr G, Ginsburg S, Herold J, Hatala R, Eva KW, Oulanova O. Using "Standardized Narratives" to Explore New Ways to Represent Faculty Opinions of Resident Performance. *Acad Med*. 2012;87(4):419-427.
22. Littlefield JH, Darosa D, Paukert J, Williams RG, Klamen DL, Schoolfield JD. Improving resident performance assessment data: numeric precision and narrative specificity. *Acad Med*. 2005;80(5):489-495.
23. van der Vleuten CPM, Schuwirth LW. Assessing professional competence: from methods to programmes. *Med Educ*. 2005;39(3):309-317.
24. Ginsburg S, van der Vleuten CPM, Lingard L. Hedging to save face: A linguistic analysis of ITER comments. *Adv Heal Sci Educ*. 2015;Online Ear.
25. Carraccio CL, Englander R. From Flexner to competencies: reflections on a decade and the journey ahead. *Acad Med*. 2013;88(8):1067-1073.
26. Donoff MG. Field notes: Assisting achievement and documenting competence. *Can Fam Physician*. 2009;55(12):1260-1262.
27. Myers KA, Zibrowski EM, Lingard L. A Mixed-Methods Analysis of Residents' Written Comments Regarding Their Clinical Supervisors. *Acad Med*. 2011;86(10 (suppl)):24.
28. Hanson JL, Rosenberg AA, Lane JL. Narrative descriptions should replace grades and numerical ratings for clinical performance in medical education in the United States. *Front Psychol*. 2013;4:668.
29. Cook DA, Kuper A, Hatala R, Ginsburg S. When assessment data are words: Validity evidence for qualitative educational assessments. *Acad Med*. 2016;(In press).
30. Ginsburg S, Gold W, Cavalcanti RB, Kurabi B, McDonald-Blumer H. Competencies "Plus": The nature of written comments on internal medicine residents' evaluation forms. *Acad Med*. 2011;86(10 Suppl):s30-s34.

CHAPTER 5

Hedging to save face: A linguistic analysis of written comments on in-training evaluation reports

Published as: Ginsburg S, van der Vleuten CPM, Lingard L. Hedging to save face: A linguistic analysis of ITER comments. *Adv Heal Sci Educ*. 2015; Online Early (DOI: 10.1007/s10459-015-9622-0)

Abstract

Written comments on residents' evaluations can be useful, yet the literature suggests that the language used by assessors is often vague and indirect. The branch of linguistics called pragmatics argues that much of our day to day language is not meant to be interpreted literally. Within pragmatics, the theory of 'politeness' suggests that non-literal language and other strategies are employed in order to 'save face'. We conducted a rigorous, in-depth analysis of a set of written in-training evaluation report (ITER) comments using Brown & Levinson's influential theory of 'politeness' to shed light on the phenomenon of vague language use in assessment. We coded text from 637 comment boxes from first year residents in internal medicine at one institution according to politeness theory. Non-literal language use was common and 'hedging', a key politeness strategy, was pervasive in comments about both high and low rated residents, suggesting that faculty may be working to 'save face' for themselves and their residents. Hedging and other politeness strategies are considered essential to smooth social functioning; their prevalence in our ITERs may reflect the difficult social context in which written assessments occur. This research raises questions regarding the 'optimal' construction of written comments by faculty.

There is growing evidence that comments written by faculty on work-based assessments such as in-training evaluation reports (ITERs) can be useful for identifying students in difficulty [1], for ranking/sorting trainees [2] and for predicting success or failure [3]. However, recent work suggests that written comments contain a prevalence of vague and 'dispositional' language [4], which faculty decode by "reading between the lines" [5]. Despite a well-established tradition of vague comments in faculty evaluation of trainees [6,7], we don't yet understand why faculty do this, how other faculty are able to decode such comments, or what their implications are for trainees [8]. A deeper understanding of these issues is an essential step towards optimizing the use of written comments for both faculty and trainees.

The presence of vague language in written comments can be a source of frustration to readers who try to interpret them. For example, comments such as "pleasant to work with" [6], or those that reflect how hard a resident worked [4], are extremely common, yet are considered particularly unhelpful for judging learners' performance – as a participant in one study noted, "if you're a good person you get those" comments [5]. One potential explanation for such comments is that faculty may not know their trainees very well, so they resort to comments that "you could say about anyone"[5]. Another explanation is that vague language is used deliberately. For example, faculty may avoid commenting on an actual deficiency to abide by the principle that "if you can't say anything nice, don't say anything at all" [5]. Another potential reason for vague comments relates to the reality that evaluation is difficult – especially when a trainee is not performing well – and that the use of vague language helps guard against negative consequences for individuals on both sides of the ITER [9].

To systematically explore the phenomenon of vague language in ITERs in more depth we turned to the branch of linguistics called pragmatics, which is concerned with how language is used for practical communication. Linguistic pragmatics argues that much of the language we use in day to day practice is not meant to be interpreted literally. Language expressing irony, sarcasm and metaphors are readily identifiable illustrations of this premise [10]. Non-literal language also includes the concept of *conventional indirectness*, [11] whereby words and phrases such as 'good' or 'meets expectations' can – by virtue of convention – come to mean below average [7]. A recent study reported many examples of non-literal language use in written ITER comments [5], but it was also found that faculty were able to rank-order residents using such comments alone with a high degree of reliability [2]. The application of linguistic frameworks to narrative assessment comments may help to explain how language that seems vague can be reliably interpreted.

Within pragmatics, the theory of politeness has particular relevance and applicability to an evaluation context. Originally developed by Brown and Levinson in the 1970's [11],

it remains influential in spite of newer theories and suggested modifications [12,13]. Brown & Levinson's framework is based on the idea of 'face', as first described in sociology [14].The concept of face has taken on different meanings over time but from Brown and Levinson's perspective, in essence, face is the public self-image that individuals try to protect. Positive face is the positive image a person has of him/herself (self-esteem); negative face is the desire to not have one's actions impeded (freedom to act). In the setting of a face threatening act (FTA), we often invoke linguistic strategies to mitigate against potential loss of face, for both the *speaker* and *hearer* (the terminology used as the theory was developed based mostly on oral language). One common example of a face threatening act is asking a colleague for a favour. It is potentially face-threatening for the person asking as there is a possibility that he may be turned down, become indebted to the hearer, or be seen as needy. It is also face threatening to the hearer, as she is imposed upon and must respond in a way that does not offend the speaker or affect her reputation as being helpful. To make such a request, the speaker may choose to compliment his colleague on something first (her recent teaching award or success at a grant competition), which redresses her positive face by enhancing her self-image and redresses his face by explaining why one might seek her assistance. More commonly we redress hearers' so-called 'negative face' by using phrases such as "I'm so sorry to bother you", or "I know how busy you are", which addresses their desire to not be interfered with. Brown and Levinson developed an explicit framework for understanding the types of language used to redress face along with the effects that these linguistic strategies have on the speaker and hearer (or writer and reader) [11]. The degree to which politeness strategies are used – and the types of strategies employed – reflect the degree to which an act is considered to be face threatening.

Hedging is one very common strategy used to mitigate against loss of face that sits within this politeness framework. Brown and Levinson define a hedge as a "word or phrase that modifies the degree of membership … in a set"; it says that the membership is "partial, or true only in certain respects" (Brown & Levinson, 1987; page 145). Other researchers have further defined and categorized hedges, including Prince, who studied hedging in physician-physician discourse [15(p93)]. Prince reported two main types of hedges: approximators and shields. Approximators affect the 'truth conditions' or a proposition in one of two ways: *adaptors* adapt a term to a non-prototypical situation (e.g., "the patient's feet were *a little bit* blue") and *rounders* indicate that a term is a rounded-off representation of a number (e.g., "the blood pressure was *about* 120 over 80"). Shields do not affect the 'truth conditions' of their propositions; rather they implicate that the speaker "is not fully committed to the belief that the relevant state of affairs actually obtains" [15(p89)]. A*ttribution shields* serve to attribute the statement to someone other than the writer, whereas *plausibility shields* introduce an element of doubt, by allowing the speaker/writer to indicate that s/he is less than fully committed

to the truth of the statement (e.g., "during my brief encounters with the resident…"). This is similar to a more recent conceptualization of hedging by Fraser, who states that hedging is a rhetorical strategy by which a speaker, using a linguistic device, can save face (for himself or others) by signalling a lack of commitment to what is said [16]. To understand hedging in our context, consider an attending's written comment that a resident's knowledge base seems "a little below average". According to Brown and Levinson, as well as to both Prince and Fraser, this statement is hedged. It could be considered an 'adaptor' (indicating the resident isn't fully in the category of below average) or it could be a 'shield' (indicating that the attending isn't fully committed to the assertion that the resident is below average). Hedges that indicate less than full commitment to the semantic category of an expression can be represented by phrases such as *sort of*, *almost*, or *like*. Another way to hedge is by not committing fully to the force of the speech being expressed, by using phrases such *as I suppose*, *perhaps*, or *I think*.

To summarize, written assessments will likely take on increasing importance in medical education, yet are often vague and can be frustrating to decode. Linguistic pragmatics – in particular, politeness theory and hedging – might help us understand and make sense of some of the vagueness in assessment language. Therefore, the purpose of the current study was to conduct a rigorous, in-depth analysis of a set of written ITER comments using the well-established linguistic theory of politeness, in order to shed light on the phenomenon of vague language in clinical assessment. These findings could ultimately be used to develop strategies to allow attendings to save/protect face while still giving meaningful feedback; and could help residents better understand how to interpret the feedback they receive.

Methods & Analysis

We compiled ITER forms for a single cohort of first year residents (PGY1's) in Internal Medicine at the University of Toronto (n=63). Each resident in this program completes an average of 9 rotations for which ITERs are generated. The attending physician at the end of every rotation completes a single ITER for the resident being assessed. The form is electronic, consisting of 18 rating scales and a single comment box; over 93% of ITERs contain comments. Our database ultimately contained comments from 1274 ITER forms. For this analysis we chose to include the highest and lowest rated residents, as extreme groups can be more "information rich" because they are unusual or differ from the norm, and can provide a useful basis for comparison. [Patton, Chapter 5] Analysis of the numeric scores indicated that less than 10% of our residents receive a 1, 2 or 3 out of 5 on any rotation so this group was chosen and labelled "low-rated"

(n=112 comments). The “high-rated” group comprised all evaluations with a score of 5 (n=525 comments).

Coding began with a line-by-line approach to each comment box (as described in more detail below) using Brown & Levinson’s politeness framework. The framework has two relevant sections for this purpose: strategies addressing positive face (by indicating that the writer wants what the reader wants) and strategies addressing negative face (which essentially “consist in assurances that the [writer] will not interfere with the addressee’s freedom of action … self-effacement, formality, and restraint” [11(p70)]. Each section has 12-15 unique strategies but not all applied to our data. Table 1 contains definitions of the strategies that were relevant to our data, along with representative quotations.

Table 1 Politeness strategies from Brown and Levinson that were found to be relevant and applicable to written comments from internal medicine residents' evaluation forms

Politeness strategies (Brown and Levinson)						
			Number of instances and percent of total comments			
Brown and Levinson strategy	Definition	Example quotations from residents' in-training evaluation reports (ITERs)	Low-rated group (ITER score = 1, 2 or 3)		High-rated group (ITER score = 5)	
Positive politeness strategies						
PP2—exaggerate interest	Exaggerate interest, approval; includes use of emphatic intensifying modifiers; absolutely, completely	Truly outstanding; excellent resident in all aspects!; an absolutely outstanding first year trainee	2	2%	165	32%
PP4—in-group identity markers	A way to claim common ground. Includes jargon/slang to evoke shared associations	Great resident; highly skilled clinician; functioning like a consultant; Dr. X is an excellent resident	40	37%	301	59%
PP10—offer	Offer, promise; demonstrates writer's good intentions	Should consider applying to X specialty; would be a great asset; will follow X's career with interest	1	1%	72	14%
PP11—presume, optimistic	Be optimistic, presume recipient will do what writer wants	Will do very well; would benefit from…; unlimited potential	26	24%	135	26%
PP15—give gifts, compliments	Symbolic gifts—thanks, compliments	A real pleasure to work with; great job; really nice to have on the team	29	27%	273	53%
Negative politeness strategies						
NP1—conventional indirectness	Use of phrases and sentences that have contextually unambiguous meanings (by virtue of conventionalization) which are different from their literal meanings	Clinical assessments are solid; demonstrated sound knowledge; functioned like a good resident	45	41%	75	15%
NP2—hedge	Word or phrase that modifies the degree of membership in a set; it is partial or true only in certain respects, or that it is more true and complete than perhaps might be expected; there are types of hedges	Is actually already quite efficient at this stage of X's training; I think X can become an outstanding clinician	103	94%	363	71%
NP7—impersonalize	Phrase as though the agent were other than the writer; avoid pronouns; delete the agent	Outstanding CCU rotation. Excellent knowledge base and clinical skills. Very mature resident	30	28%	147	29%
Bald, on record						
	No politeness strategies		76	70%	288	56%

PP refers to positive politeness and NP to negative politeness; the numbers (e.g., NP2) refer directly to the numbered strategy list from Brown and Levinson

''X'' is a placeholder used to anonymize physicians' or reisdents' names

On iterative reading and analysis we discovered that Hedging was pervasive so we coded it further by using the more detailed conceptualization proposed by Prince et al. (1982) that was briefly mentioned in the introduction. According to the authors, hedging is commonly expressed in the form of 'shields', which have two subsets.

> *Attribution shields* serve to attribute the statement to someone other than the writer. They imply that the statement conveyed is to be attributed to someone else, sometimes specified and sometimes not. Common examples are phrases such as, "the housestaff really appreciated A's teaching", or "I received numerous comments about B's performance". Statements such as "Clearly making an effort" or "Obviously has excellent communication skills" are also attribution shields as they imply that anyone – or everyone – would come to the same conclusion and thus these statements are not necessarily about the writer's own beliefs. That is, the writer's "own degree of commitment to the statement is only indirectly inferable" from what is written. [15]
>
> *Plausibility shields* introduce an element of doubt by allowing the speaker/writer to indicate that s/he is less than fully committed to the truth of the statement. They are called Plausibility shields because the speaker is making an assertion based on plausible reasons. Common examples are phrases such as, "I believe", "I think", "it is possible", "right now", etc. Statements that are marked by notation of stage of training or time of year may also be considered plausibility shields as the writer is – consciously or not – drawing our attention to these issues as a plausible basis on which to interpret their comments.

Primary coding was done by SG, who discussed the emerging understanding and application of the framework with LL. SG and LL worked together to challenge, expand and refine their understanding of the codes' definitions as they apply to our narrative data using examples from the literature along with specific exemplars from our data. Because the comments were anonymized, we could not determine the writers' intentions behind their language use. In keeping with other researchers, we therefore applied the frameworks with certain assumptions: that the comments were meant to be sincere and that the language used in this context would be interpretable in the same way as other written or spoken text. Once the coding was complete a research assistant was trained in the codes and their use, and coded sections of the data independently to assess inter-rater agreement, which was between 88 and 92% for all codes assessed. All authors then debated the interpretation of subsets of the data iteratively and critically until a cohesive interpretation could be agreed upon.

Coding was conducted without knowledge of the numeric score associated with a particular comment box. After coding was complete we were able to run a query function within NVivo to look for patterns in the coding of each performance subset (i.e., low-rated and high-rated residents). While sociolinguists don't usually use or report frequency counts in their analyses, we include them here because in applied settings they are often reported as a way to illustrate general patterns. Chi-squared analyses were used to compare proportions across group when the potential for differences was important for interpretation.

Reflexivity

The primary author, SG, is an internist and education researcher at the same institution from which the data were gathered and thus had deep insider knowledge of these narrative ITER comments (both as a researcher and as a writer and reader of the comments). SG conducted the primary coding in consultation with LL, who is a researcher in health professions education with a focus on communication. LL's background (a PhD in English rhetoric) provided a unique perspective of how language is used and interpreted in a health professions context. KE is a researcher in health professions education with a focus on judgment and assessment. His formal training led to a PhD in cognitive psychology. CvdV is an educationalist and education researcher with wide expertise in relation to assessment. KE and CvdV each questioned and challenged the interpretation and application of the codes and provided critical insights and critiques during the coding and writing process.

Results

Several elements of the politeness framework from Brown and Levinson were easily applicable to our data. As the framework is quite detailed, with much of it applying to spoken language, not every politeness strategy was relevant. We will present some of the more common strategies found in order to give a sense of the overall politeness contained in the ITER comments, after which we will delve in more detail into Hedging. Further examples and definitions of politeness strategies can be found in Table 1.

Strategies to address positive face

The most common strategy used to address positive face, according to Brown and Levinson, is called "exaggerate interest" [11(p104)], by which the writer uses emphatic intensifying modifiers to exaggerate their interest and approval. This included phrases such as "Absolutely outstanding", "superstar" and "extremely thorough and meticulous" . In addition, instances in which writers included exclamation marks were all coded here, such as, "Excellent resident in all respects!" Although the term 'exaggerate' may imply that the attending is trying to make the resident seem better than they

were, it is possible that the extremes seen in this sort of language may actually be sincere reflections of an attending's opinion. Exaggerated interest was seen in about a third of the high-rated group but only in 2% of the low-rated.

A second strategy is to use "in-group identity markers" as a way to claim common ground between the writer and recipient. [11(p107)] Phrases that include the word resident, trainee, clinician, consultant or doctor, or the honorific "Dr." were included here. Although in-group markers can be expected given the context of our data (evaluation of residents) it is interesting to note that these terms were used more often in the high-rated group compared to the low-rated (59% vs. 37%; χ^2=17, p<0.001). A third common strategy is to "Give gifts or compliments"[11(p129)], albeit symbolically, by writing that a resident is "a real pleasure to work with", or did "a great job" or was "well-liked by the team". Again these were more common in the high-rated group than the low-rated (53% vs 27%; χ^2=25, p<0.001).

Strategies to address negative face

One linguistic strategy commonly used to address negative face is called "conventional indirectness", which is the use of words or phrases that, by virtue of convention, have come to take on unambiguous meanings that are different from their literal meanings [11(p132)]. A classic example in the health professional education context is the use of the word "good", which is understood to be a code word for "below average" [7]. Other conventionally indirect phrases include "solid" and "met expectations". These strategies were more common in the low-rated than the high-rated group (41% vs 15%; χ^2=38, p<0.001).

A second common strategy, reported by Brown and Levinson, is to "impersonalize" by leaving out names and pronouns to distance the writer from the assertions made or from the recipient [11(p190)]. Consider the following example:

> Very competent team leader and team player. Looked up to by junior housestaff. Great teacher. Very professional and respectful of patients, families, and other health professionals. Very hard working and willing to spend a lot of time with patients/families ensuring issues are addressed. Thorough assessments and discharge plans.

We saw language like this in about 30% of comment boxes in both groups. The impersonalizing nature of this text might be better appreciated in contrast to the more personal language offered by a different attending about the same trainee:

> It was a pleasure working with Dr. XXXX. Although I only interacted with him for about two weeks during this month of his rotation, he impressed me with his thoroughness and interpersonal skills.

Hedging

The phrase "Although I only interacted with him for about two weeks" also offers a first look at the most common linguistic strategy we found in our data: Hedging, which was pervasive in our data, being present in 94% of comments from low-performing residents and in 71% of comments from high-performing residents (χ^2=27, p<0.001). Some examples of general hedging include phrases such as "could have been a little more rapid in starting the clinic/picking up charts to get going" or "works well, fairly independently". The words "could have", "a little more", and "fairly" are hedges because they affect either the 'truth condition' of the statement or the writer's commitment to the assertions made – in the first instance the attending could instead have written "should have been more rapid in starting the clinic" which would leave no room for doubt. Most hedges were further classifiable as either Attribution or Plausibility Shields. Table 2 contains further definitions and examples of subtypes of hedging including Approximators and Shields.

Table 2 Definitions and representative examples of Approximators and Shields that were found in written comments from internal medicine residents' evaluation forms

Approximators and Shields from prince framework (subset of hedges)						
Shield	Definition	Example quotations from residents' in-training evaluation reports (ITERs)	Number of instances and percent of total comments			
			Low-rated group (ITER score = 1, 2 or 3)		High-rated group (ITER score = 5)	
Attribution shields	Attributes belief in question to someone other than the writer	Presumably, as far as anyone can tell, apparently, clearly, noted, documented	42	39%	118	23%
Plausibility shield s (aggregated total)	Relates to doubt; writer is not fully committed to the belief	I think/feel/believe, right now, probably, perhaps	73	67%	225	44%
Subset of comments 'marked' by level, stage, time or comparison to peers			45	41%	188	37%
	Reference to level or stage of training	''Excellent knowledge base for level of training''	33	30%	129	25%
	Reference to time of year	''An excellent start to training'' ''Superb first PGY2 rotation''	8	7%	31	6%
	Norm referenced to peers	''On par with peers'' ''Ahead of peers'' ''One of the best seniors I have worked with''	10	9%	48	9%
Adaptors	Modifies or affects truth condition of proposition	Sort of, somewhat, a little, quite, really, almost	17	16%	28	5%

Attribution shields, which attribute statements to someone other than the writer, were more common in the low-rated than the high-rated group (39% vs. 23%; χ^2=12, p<0.001) and included instances in which attribution was explicit (e.g., "looked up to by junior housestaff", or "comments from multiple staff suggest …") or implicit, such as "no concerns" or "no weaknesses were identified", without specifying by whom. Often the attribution was shared, e.g., "We are all confident that XXXX will be an excellent consultant", or "Felt by all to be an outstanding resident". Attribution shields thus serve to protect the writer by obscuring his or her own contribution to the assertion.

Plausibility shields, which introduce an element of doubt, were more common, being present in 67% of comments from the low-rated group and 44% of the high-rated group comments (χ^2=20, p<0.001). Many of these comments included phrases such as "I believe", or "I think", which indicate that the writer is basing the assertions that follow on plausible reasons – their own beliefs and observations. Consider the following comment

> As far as I have been able to observe, and from the feed-forward I received from [others], XXXX has a very good knowledge base, and is a very dependable senior resident.

By using the opening phrase "as far as I have been able to observe", the writer hedges what they are about to state by indicating that they are not making claims to know anything beyond what they've seen. This protects the writer's and the recipient's face, by leaving it open to legitimate disagreement or critique by others, who may have come to different conclusions about that resident (perhaps based on different observations). The writer can't really be "wrong" because she used a hedge to create plausible doubt about what she asserted. Note that this comment also includes an attribution shield (the feed-forward received from others).

Many of the plausibility shields were marked by language denoting a resident's stage of training or the time of year. For example, "Has an excellent knowledge base for his level of training", or "good judgment for level of training", or "as good as you can perform at the PGY2 level". These statements are plausibility shields because they are qualified, or 'marked', by noting the resident's stage of training, and thus serve as plausible reasons for the assertion. Similar statements were made about the time of the year, such as "excellent first month of residency", or "excellent start". These writers may be implying – perhaps unconsciously – that their comments are meant to be taken cautiously, and are leaving themselves open to the possibility that things will change as the year progresses.

Comments without politeness strategies

By contrast to the many politeness strategies described above, we also found many examples of comments that are "Bald, on record", meaning they include no politeness language at all [11(p94)]. Examples include

> XXXX still needs to work on time management, and clinical judgement before he will be ready to take on the role of senior resident.
>
> or
>
> XXXX manages her time and the team very effectively. She is eager to learn and to teach others. She is very responsible, reliable and thorough in her management of patients.

Although this language could be found scattered throughout the comments it was rare for an entire comment box to contain only "bald" statements (12 in total).

Discussion

The purpose of this study was to use concepts and frameworks from linguistic pragmatics to gain a deeper understanding of the nature of language used on ITER comments. We found that politeness strategies were pervasive in these comments, especially hedging, which was present in nearly all of the comments about low-rated residents. We were surprised to see that it was also very common in high-rated residents. These findings support the notion that writing ITER comments is a face-threatening act, but not just for the recipient – the recurrent use of hedging in the form of 'shields' suggests that writers were also protecting their own face.

It is interesting to consider why it is face-threatening for an attending to provide written comments about residents. Brown and Levinson developed a formula for calculating the weightiness of a FTA: Weightiness = D + P + R, where D is the social distance between the speaker and hearer (a symmetrical relationship) and P is a measure of the power that the hearer has over the speaker (an asymmetrical relationship)[11(p76)]. R is the 'rank' or degree of imposition of the act in a particular context or culture. From this formula one might assume that since the attending/speaker is in the superior position, the resident/hearer doesn't have much power and the weight of the FTA is low. But consider the following example from Brown and Levinson: a bank manager meeting with an employee who wants a raise versus the same meeting but this time the employee is holding a gun – the power differential suddenly is flipped in favour of the hearer and the threat to face (and its consequences) are now very high, even with the same social distance and rank of the request. Although this is an extreme example, it

illustrates the point that the power differential is an important consideration. And in the ITER system, residents do have important power because their assessments of their teachers carry significant consequences, including the ability to affect promotion, future educational and supervisory opportunities, and the calculation of financial rewards or penalties. This suggests that our faculty are often in a position of conflict of interest and must tread carefully when giving constructive feedback.

Hedging and other politeness strategies also pervade regular day-to-day communication. Human beings are social creatures and politeness helps build and maintain relationships. As Eelen explains in an analysis and integration of theories of politeness, Brown and Levinson consider politeness to be "fundamental to the very structure of social life and society, in that it constitutes the expression of social relationships and provides a verbal way to relieve the interpersonal tension arising from communicative intentions that conflict with social needs and statuses". [17] Given the multiple roles and relationships we have with our trainees – teacher, coach, mentor, assessor, judge – and the multiple, simultaneous purposes of ITERs in the first place, we can easily envision the kinds of conflicts that Eelen describes. It should not be surprising then to see that a little politeness can go a long way when delivering feedback, and that hedging is particularly common when faculty must convey a less positive message. The use of attribution shields can allow the speaker/writer to evade responsibility for their statements, especially if there is some element of bad news. This motive of evasion may be responsible for the near ubiquitous use of hedging when commenting on low-rated residents. However, this doesn't explain why it would be so common in high-rated residents. One potential explanation for hedging in this group relates to the notion that politeness is fundamental to social life for ensuring smooth, harmonious relationships. Considered in this light, hedging should not be deemed as fundamentally problematic as it serves an essential social function.

The explanations in the preceding paragraph assume that the main recipient of the written comments is the resident, yet we know that the ITER serves multiple purposes for multiple audiences, including program directors and other attendings. The face-threat created by these other audiences may be better understood by considering the pervasiveness of hedging in other situations. For example, some linguists have studied hedging in scientific discourse and have found that it is the norm especially in the research literature [18,19]. For example, one rarely sees authors claiming to have found "the answer" to a vexing problem or "the cause" of a disease. Instead, researchers are far more likely to state that they "have found evidence that suggests that A, B or C may be factors that could be responsible for X, Y or Z". The phrases "evidence that suggests", "may be" and "could be" are plausibility shields. One reason for this sort of hedging is to protect the face of the author in case his or her claims are subsequently discredited or not reproducible. It also protects the face of other researchers who may have pub-

lished differing results or opinions. This strategy also takes into account that there may be many different (and often unknown) readers with differing levels of expertise and seniority; hedging, as a negative politeness strategy, pays deference to a broad and diverse audience, by portraying the writer as humble. The same logic applies to ITER comments – writers (i.e., attending physicians) assume there will be different types of readers (residents, program directors, competency or appeals committees, etc.), with different levels of expertise and knowledge relative to the writer. It is also quite possible that an attending might be found to be an outlier, or erroneous in their opinion of a resident, having rated them more or less highly than other attendings. By hedging – especially by using attribution and plausibility shields – the writer implies their awareness that they may turn out to be wrong and that their comments are opinions based on plausible evidence.

What are the educational implications of hedging for trainees? We don't yet know how residents read or interpret these comments but, based on the theoretical frameworks presented here, it is possible that politeness strategies obscure the intended message. The more indirect language is, the more likely it is to be misunderstood [20]. Indeed, one reason that we use politeness language is to avoid directness and create interpretive flexibility. It is also possible that residents, as savvy social members of the ITER context, are able to discern and decode the hedging and other linguistic politeness strategies in ITER comments. If so, these acts will retain their face-threatening quality despite the politeness strategies and, furthermore, residents may read hedging as an indication that there is something 'worse' that is not being said. As Brown and Levinson explain, if a writer uses a strategy appropriate to a high-risk FTA when the FTA is actually not high risk, the reader will assume the FTA was greater than it was [11(p74)]. Therefore, by using the wrong strategy or too much politeness language, we may give the impression that things are worse for the recipient than we really intend. Another issue to consider is that for many of our residents English is not the native language, and there is literature suggesting that non-literal language can be particularly prone to misunderstanding [21]. Despite these concerns, an intriguing line of research in computer-based tutoring suggests that students may actually learn *more* when language is deliberately polite than when it is direct [22]. This 'politeness effect' may depend on whether students are novices or more advanced [23] and may be worthy of further exploration in medical education contexts such as on-line learning and simulation settings.

With respect to limitations, it is worth noting that the anonymized nature of our dataset precludes us from exploring differences in language use by individual attendings – which, anecdotally at least, may be substantial. This is important as some of the politeness strategies we found, such as impersonalization, may be a reflection of the person writing, or even of the format of the ITER (i.e., a comment box at the end of a

long on-line form with 7 screens of data may lead to brief statements). To avoid over-inclusion, we did not include most of the impersonalized comments as attribution shields, as others have done; rather, we only included those that contained additional features of attribution shields. This effort to be conservative methodologically highlights a limitation described by Fraser, which is that no absolute list of hedges is possible. Nearly anything can be used as a hedge depending on how it is intended and interpreted [16]. Interpretation depends on the context, the semantic meaning, the particular hedge used, and the belief system of the recipient. Because we used a pre-existing and anonymized data set for this study, we are unable to tease apart these issues and nor can we make claims regarding intentionality. That is, we cannot know for certain what a particular attending intended by any particular comment, and our analyses have, by necessity, required interpretation of these decontextualized comments. However, we do know from previous research that clinical supervisors have a strong desire to be nice and to not offend anyone with their comments [5,9]. Based on this evidence, we have assumed that at least some of the language was intentionally polite. Further, as Mills points out in her critique of politeness theories, "politeness spans the full range from deliberate, conscious linguistic choices to the unconscious application of rules or scripts"[13(p74)]. Rather than always being a deliberate choice, some politeness language may be "determined by conformity to the norms associated with the [given] context." [13(p67)] In our context it may be that some of this language choice is relatively unconscious, which might suggest that our culture of assessment in general prefers and promotes polite, hedging language, just as our culture of scientific publishing does.

Our discussion regarding the positive and essential social functions of hedging and other politeness strategies in written feedback may serve as the basis of a new conversation in the faculty development literature on comment writing. As Watling noted, faculty development efforts to improve the ITER process have been disappointing, in part because they focus largely on enhancing faculty skill in delivering feedback and they neglect residents' receptivity to it. [8] For example, Dudek et al's work on evaluating the quality of completed ITERs reflects their faculty participants' perceptions of how "excellent" completed ITERs should look [24]. Interestingly, seven of the nine items on their ITER quality checklist focus on the written comments rather than the numeric scores, strongly suggesting that faculty supervisors feel the comments are a critically important part of the form. Efforts have arisen to train faculty to write 'better' comments [25], which focus closely on the content of the feedback, including instructions for faculty to provide "balanced" comments that include both strengths and weaknesses, to provide feedback in "a supportive manner" and to document the trainee's response to feedback or remediation. Unfortunately, training faculty has not had the desired effect [25]. Our findings may help to explain why. First, "balanced" comments are often interpreted as signalling relatively poor resident performance [5], perhaps because

concepts of assessment and feedback are conflated on an ITER, so that good "feedback" practice may not fully apply in this context. Second, the attempt to provide written feedback in "a supportive manner" may, paradoxically, lead supervisors to include *more* hedging and indirect language as a way to appear less critical. Such hedging, as we've suggested in our study, may be perceived by other faculty as damning by faint praise. That current approaches to improve faculty's written comments on ITERs may, paradoxically, distort their messages should be acknowledged and considered carefully. At a minimum, our results reveal that it is not entirely clear what (if anything) needs to be "fixed" in written comments, let alone how to fix it.

Politeness theory reveals important social motives underlying faculty's use of vague and seemingly unhelpful language when writing narrative assessment comments. Strategies such as hedging are used pervasively in low-rated residents and with surprising frequency in high-rated residents as well. This suggests that faculty attendings may be working to 'save face' for themselves as well as for their residents in the difficult social context of assessment. The social function of language in general and politeness in particular are essential and important and should not necessarily be viewed as something in need of remediation. Asking attendings to be more direct in their language may have unintended adverse consequences which should be considered in faculty development initiatives to 'improve' comment writing. Our findings also highlight the importance of exploring residents' perceptions of these hedged comments to understand how they interpret these language cues.

References

1. Cohen G, Blumberg P, Ryan N, Sullivan P. Do final grades reflect written qualitative evaluations of student performance? *Teach Learn Med*. 1993;5(1):10-15.
2. Ginsburg S, Eva KW, Regehr G. Do in-training evaluation reports deserve their bad reputations? A study of the reliability and predictive ability of ITER scores and narrative comments. *Acad Med*. 2013;88(10):1539-1544.
3. Guerrasio J, Cumbler E, Trosterman A, Wald H, Brandenburg S, Aagaard EM. Determining need for remediation through postrotation evaluations. *J Grad Med Educ*. 2012;4(1):47-51.
4. Ginsburg S, Gold W, Cavalcanti RB, Kurabi B, McDonald-Blumer H. Competencies "Plus": The nature of written comments on internal medicine residents' evaluation forms. *Acad Med*. 2011;86(10 Suppl):s30-s34.
5. Ginsburg S, Regehr G, Lingard L, Eva KW. Reading Between the Lines: Faculty's Interpretations of Narrative Evaluation Comments. *Med Educ*. 2015;49(2):296-306.
6. Lye PS, Biernat KA, Bragg DS, Simpson DE. A Pleasure to Work With: An Analysis of Written Comments on Student Evaluations. *Ambul Pediatr*. 2001;1(3):128-131.
7. Kiefer CS, Colletti JE, Bellolio MF, et al. The "Good" Dean's Letter. *Acad Med*. 2010;85(11):1705-1708.
8. Watling CJ, Kenyon CF, Zibrowski EM, et al. Rules of engagement: residents' perceptions of the in-training evaluation process. *Acad Med*. 2008;83(10 Suppl):S97-S100.
9. Ilott I, Murphy R. Feelings and failing in professional training: The assessor's dilemma. *Assess Eval High Educ*. 1997;22(3):307-316.
10. Akmajian A, Demers R, Farmer A, Harnish R. Ch. 9 Pragmatics. In: Akmajian A, Demers R, Farmer A, Harnish R, eds. *Linguistics. An Introduction to Language and Communication*. Sixth. Cambridge, MA: MIT Press; 2010:363-418.
11. Brown P, Levinson SC. *Politeness : Some Universals in Language Usage*. (Levinson SC, ed.). New York: Cambridge University Press; 1987.
12. Fraser B. Perspectives on Politeness. *J Pragmat*. 1990;14:219-236.
13. Mills S. Chapter 2. Theorising politeness. In: Mills S, ed. *Gender and Politeness*. Vol 1. Cambridge, UK: Cambridge University Press; 2003:57-120.
14. Bakker J. Facework. In: *Blackwell Encyclopedia of Sociology*. Blackwell Publishing; 2007.
15. Prince E, Frader J, Bosk C. On hedging in physician-physician discourse. In: *Linguistics and the Professions: Proceedings of the Second Annual Delaware Symposium on Language Studies*. ; 1982:83-97.
16. Fraser B. Pragmatic Competence: The Case of Hedging. In: Kaltenbock G, Mihatsch W, Schneider S, eds. *New Approaches to Hedging*. 1st ed. Bingley, UK: Emerald Group Publishing, Ltd; 2010:15-34.
17. Eelen G. *A Critique of Politeness Theory. Vol 1*. 2nd ed. Oxon, UK: Routledge; 2014.

18. Myers G. The Pragmatics of Politeness in Scientific Articles. *Appl Linguist*. 1989;10(1):1-35.
19. Salager-Meyer F. Hedges and textual communicative function in medical English written discourse. *English Specif Purp*. 1994;13(2):149-170.
20. Bonnefon J-F, Feeney A, De Neys W. The Risk of Polite Misunderstandings. *Curr Dir Psychol Sci*. 2011;20(5):321-324.
21. Danesi M. Metaphorical competence in second language acquisition and second language teaching: The neglected dimension. In: Alatis JE, ed. *Georgetown University Round Table on Languages and Linguistics 1992: Language, Communication and Social Meaning*. Washington, DC: Georgetown University Press; 1993:489-500.
22. Wang N, Johnson WL, Mayer RE, Rizzo P, Shaw E, Collins H. The politeness effect: Pedagogical agents and learning outcomes. *Int J Hum Comput Stud*. 2008;66(2):98-112.
23. McLaren BM, DeLeeuw KE, Mayer RE. A politeness effect in learning with web-based intelligent tutors. *Int J Hum Comput Stud*. 2011;69(1-2):70-79.
24. Dudek NL, Marks MB, Wood TJ, Lee AC. Assessing the quality of supervisors' completed clinical evaluation reports. *Med Educ*. 2008;42(8):816-822.
25. Dudek NL, Marks M, Bandiera G, White J, Wood TJ. Quality in-training evaluation reports--does feedback drive faculty performance? *Acad Med*. 2013;88(8):1129-1134.

CHAPTER 6

Cracking the Code: Residents' Perceptions of Written Assessment Comments

Under peer review

Abstract

Objectives

Interest is growing in the use of qualitative data for assessment. Written comments on residents' in-training evaluation reports (ITERs) can be reliably rank-ordered by faculty, who are adept at interpreting these narratives. Yet if residents don't interpret assessment comments in the same way a valuable educational opportunity may be lost. Our purpose was to explore residents' interpretations of written assessment comments using a mixed methods approach.

Methods

Twelve PGY2s in Internal Medicine (IM) participated and were asked to rank-order a set of unknown PGY1 residents from a previous year in IM based solely on their ITER comments. Each PGY1 was ranked by 4 PGY2s and generalizability theory was used to assess inter-rater reliability. Each PGY2 was then interviewed about their rank-ordering process, how they made sense of the comments and how they viewed ITERs in general. Interviews were analyzed using constructivist grounded theory.

Results

Across 4 residents the G coefficient was 0.84, and for a single resident G=0.56. Resident rankings correlated extremely well with faculty's at r=0.90. Residents appeared to be equally adept at reading between the lines to construct meaning from the comments, using language cues similarly reported for faculty. In the interviews they discussed the difficulties interpreting vague language and provided perspectives on why they think it occurs (time, discomfort, memorability and the permanency of written records) and emphasized the importance of face-to-face discussions; the relative value of comments over scores, the staff-dependent nature of assessment and the perceived purpose and value of ITERs. They saw particular value in the opportunity to review an aggregated set of comments.

Conclusions

Our findings add to the growing evidence supporting the use of narrative comments and subjectivity in assessment. Residents understood the hidden code in assessment language, perceived value in seeing aggregate comments, and accepted staff-dependent variability as a reality.

Assessment in medical education is evolving. Although most assessment is still based on numeric scores, there is increasing interest in the value of written comments, especially in residency education. [1–4] Written comments may be better than scores at identifying learners in difficulty[5] and have the advantage of being able to describe to residents both their strengths and areas that require improvement.[6] Although narrative comments can be useful, however, they are also commonly critiqued for containing language that is frustratingly vague and lacking in meaning.[7,8]

Despite this critique, recent research suggests that it is possible to use comments in a reliable way. For example, one study reported that faculty were reliably able to rank-order first-year residents (also called PGY1s, for postgraduate year 1) in Internal Medicine (IM) based on comments alone. [9,10] Another found that comments in PGY1 could predict performance in PGY3.[11] In a further study, it was found that faculty were adept at "reading between the lines" to construct meaning from otherwise vague-appearing language and that they were able to make sense of what their colleagues were expressing in their narratives.[11] The high reliabilities seen in these studies provide some validity evidence for the use of narrative comments in the assessment of residents;[12] however, the potential for comments to be useful to (i.e., to have consequential validity for) residents is unknown, because we do not know whether residents "decode" written comments in the same way as their faculty.

In one study investigating residents' perceptions of their ward assessments, Watling et al. reported widespread dissatisfaction.[13] Many residents – from multiple specialties – viewed the in-training evaluation report, or ITER, as being of "limited importance to their professional development".[13] The dominant theme identified in these interviews was the importance of engagement, on the part of both the faculty and of the resident him/herself. Similar findings have been reported by others, in obstetrics training[14] and in clinical clerkships.[15,16] These studies did not focus solely on the comments, considering broader aspects of the assessment process, but they suggest that learners may not perceive a great deal of educational value in their assessments in general. To our knowledge, no studies have explored residents' understanding or perceptions of the written comment portion of these assessments. This is important because it may help us to understand why residents perceive little value, which may in turn offer valuable guidance for faculty and instrument development. Further, if residents are misinterpreting what their attendings are trying to convey, or discounting or devaluing the messages the comments contain, then aspects of validity are at stake and formative advantages that might be gained will be suboptimal.

To address this gap, our study sought to explore residents' interpretations of ITER comments. How do they make sense of what is written? Can they rank-order residents as reliably as faculty based on comments alone? What do they see as the purpose of ITERs, and do they feel they are useful? Towards these goals, our study employed a mixed-method methodology including a quantitative, experimental component as well as interviews conducted and analyzed using a constructivist grounded theory approach.

Methods

The protocol used here replicates a study design used previously for faculty participants.[17] As in the faculty study, the materials consisted of all ITER comments collected over an entire academic year for 48 PGY1 residents in IM who graduated in 2011. There were 56 residents in that year, 48 of whom had 8 or more ITERs with comments. A document was created for each resident, including their year's worth of comments. These documents were placed into 12 packages of 15-16 residents each, such that each resident appeared in four packages and no two packages were the same. We recruited PGY2 residents from our program (n=74) as participants by generating a randomized list and sending personalized invitations in batches of 8-10 until we obtained 12 who consented (which occurred after 36 invitations were sent). We chose PGY2s to ensure that they each had a full year of experience receiving ITERs in our system. This sample size was chosen because it enabled each PGY1's comments to be rank-ordered by four participants.

Each participant was given their package of 15-16 PGY1's ITER comments, fully blinded and anonymized. They were asked to sort them into previously defined categories (A = outstanding, excellent, exemplary; B = solid, safe, may need some fine tuning; C = borderline, bare minimum, remediable; D = unsafe, unacceptable, multiple deficits)[9] after which they were asked to rank order within each category from best to worst. Following this task participants were interviewed by a research assistant (RA) who had experience with the protocol and who was unknown to the residents. Participants were asked to talk through how they made their rank-order decisions, how they interpreted the ITER comments, and to provide thoughts about the ITERs and assessment in general.

Analysis

Rank-order data were analyzed for reliability using urGENOVA with the data and study design being entered using G-string, which is a software application that allows for straightforward coding of nested and unbalanced studies. We used an all random, one-facet design with judges (our resident participants) nested within PGY1 resident being ranked. Pearson correlations were calculated (using SPSS) between the rank-orders generated by resident judges in this study and rank-orders generated in the previous study of faculty using the same dataset[17] The interviews were analyzed using a constructivist grounded theory approach, which allowed us to take an open, inductive approach to coding while being mindful of findings from similar research.[18] Coding began with consideration of a previously developed framework,[11] and sensitizing concepts included issues around subjectivity and fairness of assessments. The PI and RA who did the interviews read and discussed all transcripts as they were completed. The PI undertook primary open coding and developed a code-book with definitions and examples; during this process a second RA read and coded a subset of the transcripts based on these initial codes. The PI and RA met frequently to discuss the application and understanding of the codes, refining, merging or deleting codes as necessary until we reached consensus. The PI and RA then reviewed the key themes in depth to identi-

fy further nuances and sub-themes. The final themes were discussed and debated with the entire research team. We returned to the transcripts repeatedly to resolve disagreements in interpretation and to respond to questions from team members regarding the themes.

Results

Reliability of rankings and correlations

The mean number of PGY2 resident rankers (represented as Judges in table 1) per PGY1 comment set was 3.97. The G coefficient illustrating the extent to which the average of all rankers used the narrative comments to consistently differentiate between residents was 0.84. For a single ranker the inter-judge reliability was G= 0.56. Variance components are shown in table 1.

Table 1. Reliability of PGY2 resident judges rank-ordering PGY1 residents based on comments, along with variance components.

Source of Variance	PGY1 Resident	PGY2 Judge nested within PGY1 Resident
Estimated Variance Component	0.054	0.042
Percentage of Total Variance	56.4	43.6
Reliability for a single judge	0.56	
Reliability based on average of four judges	0.84	

Because the PGY1s forming the dataset in this study were previously rank-ordered by faculty[17], we were able to calculate correlations between faculty and resident rankings of the same PGY1s. This correlation was extremely high at r=.90 ($p<0.001$), indicating that faculty and residents ranked the same residents in largely the same order.

Interpretation of interviews

During the interviews residents spoke thoughtfully regarding their perceptions of ITER comments, and their responses touched on all of the themes previously reported from studies with faculty. Like faculty, our resident participants also "read between the lines" and took pains to explain how they interpreted what they read:

> Um, so if they used—so if they used, "excellent" specifically, like, they use specific words, I felt like it was very obvious what they were trying to say but there was a little bit of reading between the lines if they said, "good" versus, "very good". (R2)

> So many of them were generic: "excellent resident", "good enthusiasm", "happy to work with them", "pleasure to work with", I mean that's easy enough to interpret that things are good but it doesn't really give me an idea of how the resident is doing. They're very nice and positive and I would say thank you if someone told me that, but it doesn't help me get better, so that's the only reason I have trouble interpreting it per se. I know what they mean but I don't know if they're not saying things. I don't know if they're avoiding saying things they want to say because they don't want to hurt the resident's feelings or what not. (R7)

Residents did not take language at face value; rather, they made interpretations and inferences. These inferences were also influenced by their own, personal experiences with ITERs and comments as well as by rumours and what is heard through the grapevine in the IM program.

> Um, I know in my personal experience with these ITERs, there are certain rotations and certain staff that are known to rank lower, and that you have to work harder to get that "exceeds expectations". (R12)

Residents articulated the strategies they used to help them decide how to rank a given PGY1. These strategies included looking for consistency, domains commented on, specificity and quantity of commentary, and context such as which attending was believed to have written the comment (as the author identity was not known to participants); on which rotation the comment was received (which was sometimes embedded in a comment) and the time of year (which was apparent from the chronological order of presentation). These themes echo what has been previously reported by faculty[11] so will not be discussed in detail. For the purposes of this paper we will focus on some unique themes that were identified that can shed light on why vague language is likely to occur in workplace-based assessments and how residents grapple with the need to crack the code. These themes focus on the presence of vague language, the relative value of scores vs. comments, staff-dependent variability in assessment and residents' perceptions of the purpose of ITERs. It should be noted that residents' comments relate to both the ITERs read during the study task and to their own personal experiences with ITER comments.

Vague language and why it occurs

Despite the high inter-ranker reliability observed, residents struggled with vague and non-specific language, as R7 explained above, and in the following:

> Most of the evaluations I found very vague in their wording as well. "Mature, sound clinical judgment" was one that came out a lot. Um, "hardworking and enthusiastic", "good knowledge base", they say that about everyone; it doesn't really say anything. Um, another common thing people would say, um, "above expectations for this level of training". I would say three quarters of the pile was above expectations and so it makes me wonder what is "ex-

> pectations"? Are the expectations too low because everyone is exceeding them? (R12)

Residents offered many potential reasons to explain why language might be vague. These rationales grouped into several subthemes, described below.

Time and Effort

One commonly offered rationale was the apparent time and effort it takes to write "good" comments. Residents repeatedly used words such as "busy", "time consuming", and noted that it "takes a dedicated effort" to provide more detailed comments. However, some also felt that this shouldn't be an excuse: "It's part of their job description, I think, to [[pause]] do this. And it's time consuming; I'm not saying it's difficult – it's easy." (R1)

Emotions and discomfort

Residents also recognized that it could be emotionally difficult for their staff to give them meaningful feedback: "Um, and it's also hard to sometimes write about weaknesses or stuff people can work on because it sort of instills conflict." (R3) Some residents felt that faculty may err on the side of "mak[ing] their residents feel good when they should be commenting on how to make them better doctors". (R7)

Not memorable

Many residents suspected that staff may not always remember their residents in great detail, especially if evaluations are delayed, thus resulting in generic comments.

> So I could imagine trying to fill these things out for people who were, again, pretty solid residents but nothing really stood out and then trying to do that maybe a couple of weeks later would be kind of difficult. And so, yeah, you probably would go back to [comments like] "very professional", "hardworking", "good knowledge"; pretty generic statements. (R10)

Apart from the time lag, they also noted that the person filling out the ITER comments may not actually know the resident well due to limited exposure.

> But at the same time—a lot of the time—you get vague, unhelpful comments. By people who have known you for a day, or have never met you before. (R12)

Permanent Record

Residents perceived that faculty are concerned about the permanency of a written record. In many instances this was described as a reluctance to record feedback in writing because this could affect residents in the future:

> And, as such, I think that even, for example, if staff had some feedback to give you, I don't think, if they're a staff that really cares about your career, I don't think they would include it there, but rather tell you verbally. (R1)

> Um, yeah, I think most negative comments would come face-to-face, um, which I think allows them to be expressed but wouldn't necessarily hurt any future applications. (R 3)

Many residents spoke about the differences between verbal and written feedback, generally favouring face-to-face interactions. For example, R12 agreed that verbal feedback is better:

> Yes, because I think you get—you can ask questions, you can seek more feedback, you can, you know go to multiple people, get multiple opinions, um, you can also—if they say you need to work on something you can talk to them about and sort of get the back-and-forth, where the ITERs [are] just written. (R12)

Others felt that it might be "easier to say stuff to people in person" (R11) and that it would be less subject to misinterpretation, as R11 explained:

> I think it's perhaps, um, the person, um writing it doesn't want things to be misinterpreted, um, that if there is a culture that—that's just not what's usually put on ITERS, um, that they wouldn't want for it to be taken the wrong way.

The notion that constructive critique is "just not what's usually put on ITERs" emphasized residents' perceptions that putting something critical on a permanent record was largely "not done". Together, these subthemes capture residents' perceptions of why vague language occurs.

Scores vs Comments

Participants were asked about the relative value or importance of the numeric scores vs comments on the ITER. Nearly all of them expressed opinions favouring comments over scores. For example, comments were seen to be more useful, more meaningful and simply "better". They felt comments were more trustworthy than scores and more useful for providing feedback than numbers. Some felt they were easier to interpret and would be better at generating reflection. Opinions on scores varied: some felt they were more reliable and objective while others felt they were just as – or even *more* – subjective than comments. Of interest, the terms subjective or subjectivity only came up a handful of times in the interviews and applied to both the scores and the comments. Numeric scores were thought to be useful to give a quick sense of where a resident is and how they compare to others. They could be seen as being more useful for external purposes, such as comparing across residents and flagging weak residents.

In many cases residents talked about complementary purposes, such as using the comments for what is not already included in the scores or using them to flesh out high or low scores. As R5 said, "I think you can't separate the two. It's artificial to do that. I think the written comments will basically provide context to the numerical rankings and vice versa, like, you need them both". Each can provide context for how to

interpret the other. That said, several residents felt that scores should be replaced by comments, as scores are useless and arbitrary.

> So if you get like a three in communication versus a five in decision making then, you know, you don't really know what that means necessarily. Does that mean you're not good at communication or does that mean you're just not as good at communication as you are at making medical decisions? (R6)

> Maybe if it was more comment-based I think it would be helpful … but right now I just – I don't think it relays what it's supposed to relay. … Just the difficulty interpreting a 3 vs a 4 vs a 5.(R2)

Staff dependent

The next major theme encompassed instances in which residents spoke of variations between attendings, using the familiar phrase it's "staff dependent". In particular they noted differences in writing style, such as the degree of vagueness or specificity or use of adjectives and "flowery language". As one said, "I think sometimes it's like, depending on the evaluator, some people are just more descriptive" (R8); another noted that you could have two people evaluating the same resident and "one person will put "great team player, hardworking" and another one will write three paragraphs about how fantastic they were. And it could just be one evaluator's, um, much less wordy …" (R12)

Residents seem to accept that these differences are to be expected, and that, as R6 explained when reflecting on written comments in general, "You have to understand the person that's evaluating, right? And I think if you have an understanding of what they usually [write] in terms of their written comments, I think you can get a better sense of what that person is trying to convey". Consider as well the following, from R5:

> Um, and obviously there's going to be variations from, uh, evaluators or physicians or supervisors who are looking at each one, and some of them may, just by their own nature, be a little bit more verbose or more generic, which actually may end up affecting the overall evaluation. Like, for example, if you have a supervisor who tends to be a bit more generic, they may actually understate how good the resident is. Um, so it really is a lot of 'luck of the draw'. (R5)

The use of the word "obviously" reflects that, from the resident's point of view, this is something that everyone knows, while the end phrase that it's the "luck of the draw" further conveys the sense that this is just the way it is. This also applied to the numeric scores, as many residents noted differences in how different evaluators score, as explained by R2:

> So some staff like to give everyone 3s, some staff like to give everyone 5s, um, so it's difficult to interpret and, um, the residents themselves have been cautioned in interpreting these, so, not to get too bummed out if you get all 3s

> from a specific staff when you're normally getting 4s and 5s, um, because it's—I think it is very staff dependent.

Purpose of ITERs

We asked residents specifically what they thought to be the purpose of the ITER, as a way to help understand their responses to the utility of comments. They thought ITERs had several potential purposes, including to help residents, to guide the program and to meet accreditation standards.

> I mean it is obviously for evaluation and tracking progress but I think it's also great in terms of feedback and that's why I think people rely on the comments more than the numbers themselves. (R6)

Many residents felt the ITERs were – or could be – very helpful, depending on how they were conducted. "So I think ITERs in general have the potential to be fantastic, and I mean, I had very similar things in medical school at U of T, and overall I find them very useful." (R12) In relation to the written comments, particularly, they also felt that specific comments were best, not just because they helped residents know where to improve but also because they felt more personal and implied that the attending "cares more about that person." (R10) Yet although vague comments could at times be frustrating to decipher, they weren't always seen negatively. Some residents felt that they may still have a purpose: "When there is useful feedback, constructive criticism, that's also really useful as well, and the rest – it's kind of nice, I'd say, just to hear you're doing a good job." (R10)

Interestingly, the opportunity to read comments over an entire year made many of them realize how useful ITERs could be in aggregate. As R11 said, "I think that, um, on any one rotation it means nothing. I think … it would be good to actually see [comments] for the entire year all together". Similarly, reflecting on his/her own experience, R12 said, "Personally I find that when I get the sort of longer term ITERs, the six month ones, I can actually see a trend of how I am doing. I find those more helpful." According to our participants, while a single rotation's comment or score may feel like an outlier, seeing a compilation of comments can give a good sense of how a resident was doing over time. This allowed residents to see the trajectory, including which areas seemed to improve and which may have remained problematic, as R10 noted:: "In this case it was maybe a couple of things in the beginning that weren't really commented on at the end so I assumed the person improved."

Discussion

There has been ongoing debate about the value of ITERs as assessment instruments.[1] Prior research found that different faculty, across multiple institutions, can interpret ITER comments reliably. This provided needed validity evidence for using ITER comments for assessment,[12] but it was unknown whether or not residents would understand this apparently "hidden code" in faculty's often vague written language. If resi-

dents misinterpret or devalue the messages conveyed in their comments, this would potentially limit the educational value of the ITER process. We found that IM residents were able to interpret the comments with a high degree of reliability and in much the same way as faculty[17], suggesting that they are able to "crack the code" extremely well. This is despite the challenges they expressed in interpreting vague comments and making inferences and interpretations to construct meaning from the language.

Our findings can support and extend those of previous researchers in several ways. The replicability of our quantitative and qualitative findings across different groups suggests that ITER comments are indeed meaningful and potentially useful sources of assessment data provided information is triangulated across multiple judgments.[11,17] Similar to other studies with learners, our participants also valued credibility and engagement, which they often expressed as "exposure" to the attending filling out the assessment.[14–16,19] Residents' recurrent discussion of the importance and value they place on verbal, face-to-face feedback discussions reflects research highlighting the importance of a dialogue between the assessor and the learner. [20,21]

One unexpected finding in our study is that residents not only recognized that assessments can be quite "staff dependent" but that they seem to accept this fairly unproblematically. This is surprising because such idiosyncrasy of assessment is often considered to be unfair, although recent work has begun to question whether a lack of consistency between assessors should be interpreted as error or rather as meaningful variability.[22] We therefore expected our resident participants to express more frustration over staff variability, or to take the opportunity to complain about unfairness, but the overall tone of their comments was more neutral. On some level, residents seemed to accept that this is the reality – faculty are different, with different writing styles, personalities and expectations – and residents seem prepared to interpret commentary in context, both in the study and in real life. When they discussed staff variability it was in a matter-of-fact way – they would mention it and then move on rather than perseverating or complaining about it.

There are several potential interpretations of this finding. One possibility is that their nonchalance stems from the nature of the task – they were not reviewing their own ITERs but rather those of depersonalized residents who graduated years ago, so the inconsistency held no personal meaning or consequence. On the other hand, many of their comments about variability did relate to their own ITER experiences, suggesting that there is at least some element of acceptance of variability in practice. Another possibility is that seeing an entire year's worth of comments allowed them to take a birds-eye view and put outliers in context, in contrast to how they usually receive evaluations which is one rotation at a time. This interpretation is supported by comments made by some residents about how much more useful their own assessments are when seen at the six-month or one-year mark.

In this regard, our findings resonate with the medical education community's emerging appreciation for expert subjectivity and collective assessment. In a paper provocatively sub-titled "Learning to love the subjective and collective", Hodges reminds us that

"subjective" should not be equated with "biased" – indeed, as summarized from Surowiecki,[23] "many fallible judgments, summed together, create value."[4] Gingerich et al[24], drawing on work from Yeates[25] and others, suggest that variability between assessors may be a result of assessors having "legitimate but different, and sometimes conflicting, interpretations of the same observations." Govaerts[26] and Van der Vleuten have suggested that assessors should not be thought of as "perfectly calibrated measurement instruments but active agents constructing judgments"[27]; in this view, different assessors are not expected to make similar judgments and variability may actually be desirable. This emerging literature reflects exciting new ways of thinking about assessment, based on solid theoretical underpinnings as well as empirical research.

It is important to acknowledge, however, that our participants' apparent acceptance (or at least tolerance) of attending variability does not mean that residents necessarily *value* it. The question of the value of variability was not asked in the interviews, nor did it arise spontaneously, so we must be cautious in our interpretation. Residents' acceptance of some variability is an intriguing finding that can be explored further in future studies. This is important not just for educators and assessment researchers but it also more closely reflects the realities of practice – not all of our patients will see us in the same way and as professionals we need to learn to accept (and learn from) "legitimate but different" opinions of ourselves by others.

Our findings have immediate, practical implications. Given that residents saw value in seeing aggregated ITER comments, it seems that a simple intervention might involve more regular opportunities for residents to view their comments en masse rather than just at the half-way point or at the end of the year. This should complement rather than replace the current practice of viewing assessments that come month by month, which allows for timely action on any deficiencies noted. Viewing of aggregate data is supported by the literature on programmatic assessment, which emphasizes the use of multiple methods, assessment for learning, and qualitative information that relies on human judgment.[28,29] Further, the apparent value that residents place on written comments – even when vague – suggests, as others have argued,[3,19] that more commentary would be welcome. In addition, the importance of a face-to-face dialogue at the time of ITER assessments cannot be overstated. Having faculty write brief assessments every one to two weeks would have the advantages of enhancing the opportunity for feedback, allow for more aggregated comments to be seen at each time point, and could go a long way to overcome the sense of disengagement that results from frequent staff turnover. Although faculty may resist the extra work, this can be mitigated by educating them about the great value residents get from them; further, if considered as formative, the task may not appear to be as threatening.

Interpretation of our findings should consider several limitations. Our participants were volunteers and their views may not reflect those of all IM residents. Likewise we can only comment on perceptions of utility in a single, large IM program. Further studies would be needed to assess transferability to other contexts. The RA who conducted the interviews was unknown to participants but [first author] is a faculty member and attending within the same department and this may have had an influence on who

chose to accept or decline the invitation to participate, and may have affected what was discussed during the interviews.

Conclusions

Our findings add to the growing evidence supporting the use of narrative comments and subjectivity in assessment.[2,4,12] Residents in this study understood the hidden code in written comments, rank-ordered trainees reliably from them, perceived value in aggregate comment data, and accepted staff-dependent variability as a reality. Future research might explore the issue of whether residents not only tolerate but also value differing opinions of their performance.

References

1. Schuwirth LW, van der Vleuten CPM. Merging views on assessment. *Med Educ* 2004;38(12):1208–10.
2. Govaerts MJB, van der Vleuten CPM. Validity in work-based assessment: expanding our horizons. *Med Educ* 2013;47(12):1164–74.
3. Hanson JL, Rosenberg AA, Lane JL. Narrative descriptions should replace grades and numerical ratings for clinical performance in medical education in the United States. *Front Psychol* 2013;4:668.
4. Hodges BD. Assessment in the post-psychometric era: Learning to love the subjective and collective. *Med Teach* 2013;35(7):564–8.
5. Cohen G, Blumberg P, Ryan N, Sullivan P. Do final grades reflect written qualitative evaluations of student performance? *Teach Learn Med* 1993;5(1):10–5.
6. Dudek NL, Marks MB, Wood TJ, Lee AC. Assessing the quality of supervisors' completed clinical evaluation reports. *Med Educ* 2008;42(8):816–22.
7. Lye PS, Biernat KA, Bragg DS, Simpson DE. A Pleasure to Work With: An Analysis of Written Comments on Student Evaluations. *Ambul Pediatr* 2001;1(3):128–31.
8. Holmes AV, Peltier CB, Hanson JL, Lopreiato JO. Writing medical student and resident performance evaluations: beyond "performed as expected". *Pediatrics* 2014;133(5):766–8.
9. Regehr G, Ginsburg S, Herold J, Hatala R, Eva KW, Oulanova O. Using "Standardized Narratives" to Explore New Ways to Represent Faculty Opinions of Resident Performance. *Acad Med* 2012;87(4):419–27.
10. Ginsburg S, van der Vleuten CPM, Lingard L. Hedging to save face: A linguistic analysis of ITER comments. *Adv Heal Sci Educ* 2015;Online Ear.
11. Ginsburg S, Regehr G, Lingard L, Eva KW. Reading Between the Lines: Faculty's Interpretations of Narrative Evaluation Comments. *Med Educ* 2015;49(2):296–306.
12. Cook DA, Kuper A, Hatala R, Ginsburg S. When assessment data are words: Validity evidence for qualitative educational assessments. *Acad Med* 2016;(In press).
13. Watling CJ, Kenyon CF, Zibrowski EM, Schulz V, Goldszmidt MA, Singh I, et al. Rules of engagement: residents' perceptions of the in-training evaluation process. *Acad Med* 2008;83(10 Suppl):S97–100.
14. Dijksterhuis MGK, Schuwirth LW, Braat DDM, Teunissen PW, Scheele F. A qualitative study on trainees' and supervisors' perceptions of assessment for learning in postgraduate medical education. *Med Teach* 2013;35(8):e1396–402.
15. Mazotti L, O'Brien B, Tong L, Hauer KE. Perceptions of evaluation in longitudinal versus traditional clerkships. *Med Educ* 2011;45(5):464–70.
16. Bates J, Konkin J, Suddards C, Dobson S, Pratt D. Student perceptions of assessment and feedback in longitudinal integrated clerkships. *Med Educ* 2013;47(4):362–74.
17. Ginsburg S, Eva KW, Regehr G. Do in-training evaluation reports deserve their bad reputations? A study of the reliability and predictive ability of ITER scores and narrative comments. *Acad Med* 2013;88(10):1539–44.

18. Charmaz K. Coding in grounded theory practice. In: Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis. London, UK: Sage Publications; 2009. page 42–71.
19. Watling CJ, Lingard L. Toward meaningful evaluation of medical trainees: the influence of participants' perceptions of the process. *Adv Heal Sci Educ* 2012;17(2):183–94.
20. Sargeant J, Mann K V, Sinclair D, van der Vleuten CPM, Metsemakers J. Understanding the influence of emotions and reflection upon multi-source feedback acceptance and use. *Adv Heal Sci Educ* 2008;13(3):275–88.
21. Sargeant J, Eva KW, Armson H, Chesluk B, Dornan T, Holmboe ES, et al. Features of assessment learners use to make informed self-assessments of clinical performance. *Med Educ* 2011;45(6):636–47.
22. Gingerich A, Regehr G, Eva KW. Rater-Based Assessments as Social Judgments: Rethinking the Etiology of Rater Errors. *Acad Med* 2011;86((10 Suppl)):S1–7.
23. Surowiecki J. The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations. 1st ed. New York: Doubleday; 2004.
24. Gingerich A, Kogan JR, Yeates P, Govaerts MJB, Holmboe E. Seeing the "black box" differently: assessor cognition from three research perspectives. *Med Educ* 2014;48(11):1055–68.
25. Yeates P, O'Neill P, Mann K V, Eva KW. Seeing the same thing differently : Mechanisms that contribute to assessor differences in directly-observed performance assessments. *Adv Heal Sci Educ* 2013;18(3):325–41.
26. Govaerts MJB, van der Vleuten CPM, Schuwirth LW, Muijtjens AM. Broadening perspectives on clinical performance assessment: rethinking the nature of in-training assessment. *Adv Heal Sci Educ* 2007;12(2):239–60.
27. van der Vleuten CPM. When I say … context specificity. *Med Educ* 2014;48(3):234–5.
28. van der Vleuten CPM, Schuwirth LW. Assessing professional competence: from methods to programmes. *Med Educ* 2005;39(3):309–17.
29. Schuwirth LW, van der Vleuten CPM. Programmatic assessment: From assessment of learning to assessment for learning. *Med Teach* 2011;33(6):478–85.

CHAPTER 7
Discussion

The studies presented in Chapters 2-6 will be briefly summarized, following which their collective findings will be synthesized in order to answer the research questions posed in Chapter 1.

To begin, the study reported in Chapter 2 found that it is possible for faculty to rank-order residents based only on the comments other faculty provided, showing that resident performance can be captured through narrative comments in a reliable way. In Chapter 3 we reported that faculty were able to do this by reading between the lines and interpreting an apparent "hidden code" in the comments, although they were somewhat frustrated by generic language. This study raised several critical questions that we then attempted to address in subsequent studies. In Chapter 4 we found that it did not matter if faculty were insiders or outsiders – reliabilities remained high even when comments from one institution were given to faculty from other institutions, suggesting a degree of universality to the "hidden code". Further, using only the first three months of comments yielded comparable reliabilities to using the entire year's worth. In Chapter 5 we used politeness theory to explore why attendings write the way they do – with a preponderance of vague, generic and dispositional language – and found that comments are constructed to allow faculty and residents to "save face". Hedging language may be crucial to enabling smooth social interactions and this may help explain why it persists. Finally, in Chapter 6 we answered the question of whether or not residents can also interpret assessment comments in a reliable way and found that they are just as good as faculty at this task. Similar to faculty, they did not take language at face value, and despite the pervasive use of non-literal language they formed similar interpretations when compared to faculty. Residents also appeared to accept some degree of variability in attendings' commenting style which may lend support to the emerging push to embrace more subjectivity in assessment.

Answers to Research Questions Posed in Chapter 1

In Chapter 1 I posed five research questions, which will be answered briefly below. The responses here are deliberately concise because, as I iteratively reviewed the findings in Chapters 2-6, I came to realize that there were broader themes and lessons learned that are best discussed not as responses to individual questions but rather by considering all studies in aggregate. This discussion will follow.

Research Question #1
Can narrative comments on in-training evaluation reports (ITERs) be used reliably to discriminate between residents in Internal Medicine?
Our findings demonstrate that comments can indeed be used reliably for this purpose. In Chapters 2, 4 and 6 we found that PGY1 residents in internal medicine can be rank-ordered based on comments alone with incredibly high reliability. This finding holds up across external faculty as well as with residents as raters.

Research Question #2
How do readers make sense of narrative comments?
As presented in Chapters 3 and 6, faculty and resident participants did not always take language at face value, but rather were adept at reading between the lines, actively using cues in the language to make sense of written comments. Chapter 5 shed further light on how non-literal language is used and interpreted, while offering indications that such language is maintained because it can be adaptive despite its often frustrating ambiguity.

Research Question #3
Can narrative comment analysis be used as a feasible approach to assessment?
From the data shown in Chapters 2 and 4, we can suggest that it probably can. It is not possible to be more definitive because, although we showed that the task can be completed with only 3 months of comments and by faculty who are not specifically trained, and also found that residents can do it just as well, we did not actually test feasibility in a practical sense. This would have to be explored in future research.

Research Question #4
Why do clinical supervisors write the way they do?
In Chapter 5 we saw that, when faculty attendings write ITER comments, a great deal of their language is hedged. The use of multiple forms of politeness strategies suggests that assessment is a face-threatening act. This could explain why so much of our assessment language is vague or appears so, and it has implications for redirecting the conventional faculty development of trying to "improve" faculty comments through less hedging. As we saw in chapters 3 and 6, readers (both faculty and residents) picked up on this hedging and on a general deficiency of direct, specific language, and saw meaning in it.

Research Question #5
Do residents themselves understand what is written in ITER comments? Do they perceive educational value in this approach?
From the study reported in Chapter 6 we learned that residents do seem to understand what is written in the ITER comments; they are as adept as faculty at reading between the lines and using language cues to actively construct meaning. Residents also seem to perceive value in the ITERs if they are done well – on time, from people with whom they interacted, and with helpful written comments. Of note, residents also seemed to accept or at least tolerate some variability between attendings, although we do not yet know if they value these variable opinions.

As I reflected on the five studies together, I came to see 4 major themes, or lessons, that draw on the work presented here as well as on the literature in relevant domains. These themes represent the Integration phase of the multiphase, mixed methods approach described in chapter 1. These 4 themes are:

1. Written ITER comments are useful and should no longer be neglected as valuable sources of assessment data. Our findings can be used to re-interpret past research and help resolve apparently discrepant evidence.
2. The language we use in assessment can appear to be frustratingly vague and generic, but there is enough meaning embedded that people can make sense of it, whether insiders, outsiders or residents. There is a hidden code but it seems that people can generally decode it, which suggests a common/shared understanding of performance.
3. We need to recognize the powerful social need to save or protect face (on the part of both faculty and residents) when faculty attendings write ITER comments. This finding may lead to better faculty development initiatives.
4. The analysis of written comments – and readers' responses to them – can lend support to continuing arguments to embrace more subjectivity in assessment.

In the following sections I elaborate on each of these themes in turn.

The utility of ITER written comments

In light of the research presented in Chapters 2-6, it seemed important to go back and critically reappraise the literature, as our findings strongly suggest that ITERs and their written comments may be more useful than what is usually reported. Articles and opinion papers from decades ago denigrated ward assessments such as ITERs as being too subjective and so fraught with bias of all sorts that they should not be used as significant sources of assessment.[1,2] Yet, as outlined in Chapter 1, evidence exists to counter at least some of these dire warnings, such as research that showed an enhanced ability to pick up learners in difficulty based on their written assessment comments with more sensitivity and at earlier time points than when based on numeric scores.[3] Even recent authors have strongly recommended giving up ITER-type assessments altogether, stating that our belief in the "face validity" of these sorts of assessments is misplaced,[4] and arguing that the preponderance of non-specific comments are "at best useless and at worst detrimental to learners' progress." [5] In this section I will explore the discrepancies between these arguments and the data reported in this thesis.

Turning our attention to the comment portion of assessments, the findings reported in this thesis may seem to contradict much of what has been written in the literature. However, if we take a closer look at some of these studies and reappraise them in light of current findings, the apparent differences may be explained. Many previous researchers, for example, have come to the conclusion that written comments are "too generic" to be of use. In a 2001 study involving medical students in pediatrics, Lye et al. were dismayed that evaluation comments were more often about learner and personal characteristics (approximately 50% of comments) when compared to comments about basic clinical skills. Yet a look at what they coded as personal "characteristics" includes such things as *respectfulness*, *professional*, *conscientious* and *seeks feedback*. Comments involving these domains were deemed to be non-useful and not specific

enough to lead to behaviour change. Similarly, Ringdahl et al. reported that generic comments and personal attributes made up about 38% of comments compared to "clinical competence" which accounted for only 14%, concluding, similar to Lye et al., that most of the comments do not "help learning".[6] In neither of these studies is it possible to determine the impact on learning; but the more important point is that although comments regarding clinical skills are surely desirable, it doesn't necessarily follow that comments about these other domains are *unimportant*. Indeed, as argued in Chapter 1, as a medical education community we are supposed to be assessing far more than clinical skills.[7] The assumption that "personal" or "generic" comments are unhelpful should thus not be automatically assumed or taken at face value.

Other, more recent studies can also be re-evaluated in this light. For example, Vivekanada-Schmidt et al. in 2013 reported their analysis of free-text comments on multisource feedback of practicing physicians in the UK and concluded that "in their current form, the overwhelming majority of free-text comments add little to facilitate improvement in assessees' personal development and performance."[8] Yet a critical look at their work, in light of current findings, may lead to a more favourable conclusion. For example, one of their biggest criticisms was that comments were more rater-centred than learner-centred, focusing on the assessee's effect on others; yet many of the examples given related to concepts like helpfulness, teamwork, and availability for support and supervision – which, as previously mentioned, are elements of competency that we are *supposed to* be assessing. Similar to the above studies, these researchers were concerned that the written comments were not behaviourally based or specific enough to be helpful for learning. However, looking at the example given of such a non-helpful comment ("Dr. R demonstrates excellent involvement of the multidisciplinary team") its lack of utility is not obvious and we don't know whether or not it was helpful for Dr. R. One final recent study is worth mentioning, as it involved an assessment of the quality of written feedback on 500 written comments on internal medicine residents' ITERs, a dataset similar to the ones used in our studies.[9] These researchers similarly concluded that most of the feedback was not helpful as nearly 50% of it was non-specific or about personality. What is unique about this study is that it highlights a critical issue that had not yet been reported: the measure used to assess "quality" was based on criteria generated by participants during small group exercises – yet despite this collaborative process there was "considerable discrepancy between groups in determining when statements were sufficiently specific."[9] It is somewhat perplexing that the major critique focused on a supposed lack of specificity, while at the same time consensus on the definition of "specific" was so problematic.

Two inter-related concepts may help further explain why our findings seem to differ from those of other researchers: feedback and specificity. Why are specific comments so desirable, as echoed repeatedly in prior studies? In the literature cited above, the focus has been on the utility of written comments as *feedback*, and it is generally accepted that feedback should be SMART: specific, measurable, achievable, relevant and targeted appropriately.[8] Most of the studies referenced simply equated written comments with feedback, and applied this framework to their interpretation. For example, Jackson et al. used as their database written comments from internal medicine resi-

dents' ITERs and called the comments "written feedback".[9] The study by Canavan et al. similarly used comments from an instrument called the "*Assessment* of Professionalism Behaviours" but framed all the comments as forms of feedback.[5] Yet we know from Watling et al. and from our own work that there are multiple perceived uses of an ITER in addition to providing feedback, including documenting progress, sending messages to the program director that there may be concerns, and serving as the basis for reference letter writing.[10,11] As discussed in Chapter 6 and as reported by Patel et al.,[12] residents and many faculty feel that the verbal, face-to-face conversation that takes place along with the ITER is perhaps more important than what is ultimately recorded in writing, as it allows for an exchange of information resulting in clarification; further, it avoids the fear expressed by faculty and residents of formal documentation on the ITERs, which are seen as a permanent record that can hurt residents in the future. Our residents also noted that sometimes "it's just nice and encouraging to hear you're doing a good job," even if that comment is not behaviourally based or specific. Therefore, arguments that non-specific comments are not useful may be overstated, as ITERs are not necessarily (or solely) about feedback; further, the noted difficulty in even deciding what "specific" means suggests that a critique based largely on the issue of specificity is problematic.

There is another aspect to specificity in written comments that is worth discussion. Our participants did express a desire for more specificity in the comments they read, but for slightly different reasons than are suggested in the literature. The more specific a comment was (e.g., if it described an event) the more credibility it had, partly because it gave the impression that the attending really knew the resident and was therefore in a position to assess their abilities. Vague and generic sounding comments were given less weight, especially if they were outliers. Specific comments also appear to signal that the attending "could be bothered" to write something in detail, which implies a level of commitment to the resident's development that a vague comment does not. Residents felt that specific comments provided more information about how to improve (the feedback function), but they were not entirely dismissive of generic comments. In fact, if generic comments were relatively vague but also fairly similar it could still give a sense that the resident was doing well overall. Taken together, it seems that while there still may be benefit to encouraging more specificity in comments, this criterion should not necessarily be of overriding importance in determining quality.

Language is vague but decodable; the "hidden code" is accessible

As noted throughout this thesis, there have been hints in the literature that the language encountered in assessment contexts is not used or interpreted in a literal sense. Excellent isn't excellent.[13,14] Good means below average.[15] Below average is really only the bottom 10%.[16] In the studies reported in Chapters 2-6 we similarly found that language was often not taken literally but meaning was inferred based on language cues, contextual factors and norms. We argued that a hidden code exists and, because it

seemed to be readily accessible to faculty and residents, we began to understand it as tacit, but not hidden after all.

Our research helps explain how otherwise vague and generic language can still be interpretable. One key methodologic choice we made was instrumental in shedding light on this phenomenon: we went beyond coding fragments of comments to determine whether or not they are specific enough, positive or negative, behaviourally based (enough) or deemed to be helpful. We looked at comments whole and in context, rather than as parsed statements. This strategy allowed readers to use contextual cues including timing and chronological order, rotation type, domains commented on, consistency, etc. The context was not always explicit and, therefore, was often inferred or guessed at, but even so, it was helpful with rank-ordering. Even though raters expressed frustration with vague language, they were quite good at deciphering it and making sense of what was written. This body of work extends research by Bogo et al, outlined in Chapter 1, which studied preceptors' assessments of social work students and reported that important context would be lost if sentences were parsed into unrelated phrases. They found that preceptors used statements with "but" when they were attempting to excuse deficits, such as "the staff liked them but they were too casual" or a student had "good practice skills … but they had problems in relating interpersonally". These statements were viewed as helpful in that they allowed the reader to see how the writer couched or explained their otherwise positive or negative comments.

Although there are likely many explanations for why this linkage is important to maintain when interpreting language, one theory from social psychology – the "faint praise effect" – appears particularly relevant. There is an expression in English that one can be "damned by faint praise", which means that praise can have a negative effect on impression formation if it is not as strong as might be expected. Put more formally, it means our beliefs about something or someone can paradoxically become more negative following the receipt of only positive evidence.[17] If we receive weak positive evidence about someone when we expect strong evidence from that person, we make inferences about why that stronger positive information is absent. For example, if an attending writes only that a resident is pleasant and hardworking we might infer that they are not clinically excellent – otherwise the attending would have said so, as such comments are thought to be particularly informative and it is within the attending's position and purview to make such statements. When this expected information is absent we make negative inferences about what isn't expressed. In contrast, if the attending says that the resident is excellent clinically, highly knowledgeable and also pleasant and hardworking, we don't make the same inference because the weak evidence is preceded by something stronger. That is, the same words – pleasant and hardworking – are interpreted very differently in the two contexts. In retrospect, this effect could help explain findings we reported in an earlier study, in which attendings' impressions of residents "did not seem to result from a linear sum of dimensions; rather, domains [that were described] idiosyncratically took on variable degrees of importance… Relative deficiencies in outstanding residents could be overlooked, whereas strengths in problematic residents could be discounted."[18] We recognized that the order in which information was discussed was important to capturing attendings' im-

pressions. The "faint praise" effect might help explain why this occurs, and it supports our arguments to use assessment comments "whole" – meaning might not only be lost but distorted if it comments are parsed.

The strategy of using whole language in our studies may be one reason why our participants were able to interpret language so effectively. Regardless of the cause, it is clear that they have a shared understanding of what performance should look like for a PGY1 in Internal Medicine. This is suggested by the reproducible rank-ordering of residents based on only written comments – if everyone can understand the "hidden code" in assessment language and weight comments in the same way, they must have some agreement as to what is most important to performance. Some researchers have suggested that a shared understanding of performance and expectations would enhance reliability and validity of assessments and indeed may be critical to their success.[19] We found in earlier studies that when faculty describe the performance of their trainees, their descriptions do not map easily onto the CanMEDs framework.[18] This lack of fit also applied to the written comments themselves[20] and suggested that some difficulty can arise from a disconnect between what faculty are asked to focus on and what they appear to focus on more naturally. These two studies created a framework to describe how faculty seemed to be conceptualizing their residents' performance, which included not just competencies, such as clinical care of patients or communication skills, but also "non-competency" domains such as disposition, attitude, and trajectory of performance. Similarly, Kennedy found that supervisors tended to consider four dimensions of their trainees that led to entrustment: knowledge, discernment of limitations, conscientiousness and truthfulness.[21] Studies like these remind us that, while we may share mental models of performance in certain contexts, these models may not fit well with official competency frameworks; instead, they may be constrained by them.

Determining what – and how – faculty think about the performance of their trainees may be a way to improve assessment. Work by Crossley et al. found that if assessment language is aligned to constructs that are relevant to the context and to how assessors conceptualize performance, we can greatly improve reliability.[22] This is not easily reconciled with our current assessment system which is "roles" based, requiring faculty to assess learners on the seven CanMEDs roles as distinct entities. About competency frameworks and ITERs it has been said that "the sum of the parts detracts from the intended whole".[23] This statement applies for both competency frameworks and for the language used in assessment. Crossley et al. suggest that we should trust the judgment of the experts who observe, teach, coach and assess residents on a daily basis.[24] Their opinions and judgments are reflected in the words they use – we just need to harness them better.

Despite their potential, the acceptance of generic, non-literal language for assessment purposes may also entail risks which should be considered. Consider, for example, the issue of non-native language speakers. In Chapter 6 we reported that residents seem to be as adept as faculty at reading between the lines and interpreting written comments. Of the 12 participants, however, only one self-identified as an international

medical graduate (IMG). Non-native language speakers are known to have difficulty in using and interpreting politeness cues, such as non-literal language,[25–27] and IMGs have been noted to have lower success rates on some high-stakes exams at the end of residency training.[28] It would be important, therefore, to determine if IMGs would interpret comments the same way as their Canadian peers.

Another potential risk is that the language attendings use may inadvertently send unintended messages to their residents. As with the "faint praise" effect described earlier, and as described extensively in Chapter 5, non-literal and indirect language can misfire. If not much is written, and if the writing is generic, it could just be that the attending was quite busy that day or that he or she didn't spend enough time with that resident to feel comfortable writing more extensively. However, the reader could form several different impressions, such as that the resident wasn't very memorable or the faculty didn't care enough about him/her to write more. The use of weak language or faint praise could be interpreted as a strategy to avoid writing something more negative. Similarly, if the politeness strategies used seem out of proportion to what is expected, it can make the interaction seem more face-threatening than it actually is.[29(pp76-81)] Even if any of these messages are intended, they are quite subject to misinterpretation by an individual learner and could have unintended negative consequences. Attendings are likely not aware of the range of potential misfires that could result from their wording; therefore, this could be an important focus of future faculty development.

The importance of saving face

One powerful finding of our research, described largely in Chapter 5, is that assessment appears to be a face threatening act, both for faculty writing the comments and for the residents receiving them. While others have noted that faculty are reluctant to give negative feedback or to fail students, the use of politeness theory allows us to enhance our understanding of why this occurs, especially in more benign situations that don't involve failure or poor performance.

Research in this domain has often focused on faculty's unwillingness to record negative evaluations, the most dramatic example being the phenomenon of "failure to fail."[30,31] In an interview study by Dudek et al., faculty participants were confident in their ability to determine whether or not a trainee was performing poorly, yet they identified four major barriers to failing such a student: lack of documentation, lack of knowledge about what to document, anticipating an appeal and lack of remediation options.[31] This has been reported in other health professions, notably by Ilott and Murphy, who also delved into the emotional responses that occupational therapy educators experience around failing students.[30] This study noted that feelings of extreme anxiety start during the decision-making phase and often lessen after the "bad news" has been given to the student, although the relief was often mixed with feelings of self-doubt and guilt. On the other hand, respondents often reported pride in fulfilling their important professional obligations and in making "the right decision for the

right reasons". These sorts of studies are very helpful to the education and practice community because they can lead to better supports and resources for faculty who have identified underperformers.

The research on failure to fail can also be interpreted as supporting the notion of assessment as a face threatening act. It is face threatening both for the learner, whose failure will pose a powerful threat to professional identity, and for the faculty, whose decision to deliver this bad news can entail a range of negative repercussions. When we applied a politeness theory framework to our residents' assessment comments we were, therefore, not surprised to find abundant evidence of linguistic strategies such as hedging and shields in our lower-rated residents' comments. What was more surprising was to find such extensive use of them in the comments associated with high-rated residents. This strongly suggests that *any* written assessment can be face threatening, not just those encounters that involve failing. Potential reasons for this were explored in Chapter 5, the most compelling of which is that "politeness is fundamental to social life for ensuring smooth, harmonious relationships."[32] We further suggested that "hedging should not be deemed fundamentally problematic as it serves an essential social function."

How can we reconcile this basic, social need for smooth social interactions with the desire for more specific and constructive assessment comments? The relationships that faculty have with their trainees are complex; faculty serve multiple roles at once, so that means that they are never "just" assessors. Perhaps part of the problem is that educators and researchers tend to equate assessment with feedback and this may be problematic. As seen in our studies and reported by others, a strong preference exists for verbal, in-person dialogues for purposes of discussing feedback, so looking to ITER comments for feedback information may be misguided.[12] Instead of focusing on what assessment comments lack, we might learn something quite different from what they contain, even when what they contain is often criticized. As explained earlier, non-specific, generic and personal comments are universally criticized as being "at best useless and at worst detrimental" to learners' progress.[5] Yet, viewed in light of politeness theory, they may be serving a critical purpose, by actively addressing these implicit social relationship needs. In some of our interviews and in a separate study we found that faculty sometimes use our ITER almost as a way to thank residents for working so hard, especially when the rotation has been very heavy.[33] Our resident participants noted as well that sometimes it's just "nice to hear that you're doing a good job".

There are important reasons why educators and researchers shouldn't neglect or negate the important social function of some types of assessment comments. One is that generic, non-specific, or weakly positive comments can pave the way to allow for more critical feedback to be accepted. The classic "feedback sandwich" relies on this phenomenon; however, although students view the approach favourably, its efficacy in improving learning is in question.[34] Eva et al. described the emotional nature of seeking and receiving feedback, including the conflicting desires within healthcare professionals to develop and improve while also not looking stupid; this was also framed as a desire to not lose face with peers, supervisors, trainees and even with oneself. From

this perspective, we may understand that it is possible that vague, weakly positive language in the "feedback sandwich" may be an assessor's attempt to cool down the emotions that trainees experience, thereby allowing face saving. Other researchers have also emphasized the importance of being attuned to recipients' sometimes intense and long-lasting emotional reactions, which can impair their ability to process and act on the feedback they receive.[35,36] These reactions are linked to learner satisfaction, which can also be negatively impacted by the way in which feedback is given. Boehler et al. reported that when students were given praise, they rated their instruction higher than if they received constructive feedback, despite the fact that they learned significantly more from the feedback than from the praise.[37] These studies together highlight the tensions between giving constructive feedback that will help learning while at the same time allowing learners – and their assessors – to save face. It is important to note that while these studies were focused on feedback and not assessment, it is likely that the same concepts would apply to assessment language, perhaps even more so given the stakes involved. It is possible that the use of nonspecific language is an attempt to resolve this tension.

Another consequence of neglecting the social purposes of language in assessment is that doing so may limit the success of our faculty development efforts. To date, research efforts to create a scoring rubric for ITER assessment comments for faculty development purposes have shown modest improvements at best.[38,39] The studies were small and it is not known whether the changes observed were sustained. More importantly, it is unknown whether the "improved" comments are more useful in distinguishing between learners, or how others would view them. For example, if residents in one program began receiving more comments indicating areas for improvement, and the culture was unchanged elsewhere, these residents might be inadvertently disadvantaged in the eyes of other program directors or attendings. We need to pay close attention to how residents themselves would view this "improved" ITER language. As Watling et al.[40], Eva et al.[41] and others have discussed in depth,[36] paying insufficient attention to how recipients react to and process their feedback (or assessment) can greatly limit educational impact. This is meant neither as a critique of these important efforts, nor as a reason to avoid moving forward; rather, it is meant as a call to consider a broader perspective. These effects and conflicting purposes require careful research to disentangle.

In summary, it seems we cannot underestimate the powerful need to save face, both for those writing the assessment comments and for those receiving them. As noted by Harris et al., as our social relationships become closer, "politeness concerns will lead to an even greater reluctance to express negative information."[17] This is part of our fundamental nature as social beings and it should not be expected to be terribly different just because individuals may also be assessors. Faculty supervisors use politeness strategies, hedging and shields in order to protect and preserve face. What our studies suggest is that this may not be as dire a problem as is usually depicted. If it is generally possible to decode the comments while saving face, then perhaps no real harm is being done.

Subjectivity

In Chapter 1 I reviewed some of the recent literature that is pushing us as a community to embrace more subjectivity and collectivity in our assessments. There were several reasons for this push, including the reality that we need to evaluate ever more competencies in a more continuous way, and the fact that tests built on psychometric models will not suit every purpose. In addition, not all competencies lend themselves well to "testing"; some need to be observed and assessed in situ in order to maintain their authenticity. This argues for maintaining some type of ITER instrument as an element of workplace-based assessment. In a recent article, Eva et al. emphasized that "authenticity is achieved when assessment protocols accurately reflect the domain of practice such that 'studying to the test' or learning to 'game the system' equates with learning to practice well."[42] Such assessments could also serve important purposes for life-long learning and for models of continuous assessment of competence. Assessments that include descriptive comments would arguably align better with this ideal than scores alone, as they give the learner a better sense of what and how to improve.

The value of multiple sampling, long recognized to be important in quantitative assessment, can also be recognized as a strength of written assessment comments. The psychometric literature strongly recommends sampling across multiple contexts and by multiple assessors as the best way to ensure reliability.[43] Further, it is now becoming accepted that many subjective opinions added together – especially if independent – can paint a more accurate picture of the whole than would be obtained from only one or two opinions.[44–46] When we think of ITERs we tend to view each one as a complete assessment in and of itself, rather than as part of a system in which an average of 10 per year are gathered. For most assessments this degree of sampling would be considered adequate, and it appears to be the case for ITERs too, both numerically and with respect to the comments.[16,47] ITERs should therefore be thought of collectively: each one is a single data point that informs a larger picture of competence for residents. Interpretation has to occur not only at the level of each individual ITER but across all ITERs for each resident. Assessment systems should be designed to allow for such integration to occur. The findings presented in Chapters 2, 3, 4 and 6 lend further support to these recommendations.

Our research with residents provides two new pieces of evidence in this regard. The first is that residents did not view the ITERs as negatively as we thought or as previously reported,[48] and they did not use terms like "too subjective" or "just subjective" as a way to dismiss them. They also did not have a consistent view of either comments or numbers being more or less subjective than the other. The second is that residents seemed to accept that attending variability exists and is not necessarily terribly problematic. As described in Chapter 6, because we didn't ask specifically whether residents value variability, we cannot make definitive statements in this regard; however, it is important to note that they at least accept it. If this is borne out in future studies – especially those that explore the extent to which they value variability – this will bode well for residents as life-long learners. The increasing requirements to accept 360° multi-source feedback, including views of patients, will necessitate that residents learn

the skills to process potentially disparate viewpoints without simply dismissing outliers. Through ITERs, novices can learn early on that subjectivity and rater variability are not in themselves bad things. Although some researchers have suggested that we identify and remove inconsistent raters (i.e., outliers)[49], or simply average out all scores, this could have the effect of discounting or obscuring potentially valuable opinions. Instead, we may do better to capitalize on residents' acceptance of variability and more explicitly teach them how to consider and use discordant opinions in context. Put another way, rater variation, usually treated as "noise", may be better thought of as diversity of "signal" which could inform learning.

There are some potential caveats to the use of subjectivity that need to be considered and addressed. No one would welcome a return to the days of pure subjectivity with insufficient sampling, as in the old "long cases" used in high-stakes exam settings.[50] Subjective assessment still needs some sort of framework and benchmarks, especially for novice attendings.[51] There is also an important distinction to make between subjective and biased – bias implies that impressions are being formed or judgments are being made based on construct-irrelevant factors, such as race, ethnicity or age. We also need to ensure that observations and assessments are truly independent; if they are not, "group think" can take over thereby removing the advantages of sampling multiple perspectives. Removing outliers may improve reliability, but we would lose the opportunity to learn from them – what are they seeing that may be different from what others see?[52] What can the learner learn from this unique perspective?

In summary, although the benefits of embracing subjectivity and collectivity are potentially significant, we must start with a clear purpose to our assessments, consider them in the context of a program, rather than as individual instruments,[53] and ensure that there is enough independent sampling. Our studies support this view and add to it the valuable resident perspective, which should enhance acceptability.

Discussion of Methodology

As described in Chapter 1, this body of work is based on what Creswell and Plano Clarke would define as a multiphase, mixed-methods approach.[54] The four themes discussed in the preceding paragraphs represent the "integration" phase of this research. Each theme drew on multiple studies and on the relevant literature in a far deeper way than what could be achieved within each individual study. Indeed, this is the strength of multiphase research. In this section I will consider what our approach allowed and enabled as well as what it might have obscured or made more difficult.

Each study was driven by a particular question and required a particular methodologic approach. The studies on reliability of ranking by comments versus scores clearly required a psychometric approach. To understand how readers were interpreting the comments in order to rank-order them, we needed to ask them why – this lent itself to grounded theory, as we were developing a framework to understand how this was done. We chose constructivist grounded theory as it allows for open, inductive coding

while attending to either previously developed frameworks or "sensitising concepts" from the literature. Both of these approaches were taken in the study described in Chapter 6, which reported both quantitative and qualitative findings together. By using constructivist grounded theory in Chapter 6, we were able to validate and extend the framework developed in Chapter 3, and by using Generalizability (G-)theory we were able to add to the psychometric evidence in Chapters 2 and 4. We used a different method to try to understand why faculty write the way they do – for this purpose we turned to linguistic pragmatics, in particular politeness theory, and used a mixed approach (qualitative analysis with numbers). Each one of these methodologic approaches was chosen for being the most suitable for the specific question posed and each was conducted according to the appropriate principles of rigour. Taken together, they allowed us to create a more complete understanding of the use and interpretation of language in assessment comments from multiple perspectives.

In terms of reliability estimates, the G-studies allowed us to determine and compare reliabilities across various groups. They also made it simple to conduct decision studies to estimate the reliabilities that would be obtained with different numbers of raters. This can greatly help in decision making when it comes to designing assessment innovations in real life settings. The D-studies we presented can allow educators to consider the number of raters that they could feasibly recruit in their settings and quickly assess the trade-offs in terms of reliability.

From the interviews we were able to develop a framework with constructivist grounded theory and then extend it to help describe and understand residents' views. This framework was instrumental to our understanding of what it was in the language that allowed readers to construct the reliable interpretations we observed. Furthermore, the initial framework in Chapter 3 triggered the study reported in Chapter 5. The interviews had given us hints as to why language might be vague and generic, which led us to politeness theory and hedging. The politeness frameworks we used were powerful tools with which to analyze written comments, and led to a new understanding of why vague language might be so prevalent. This in turn – along with the other studies conducted – led us to the inevitable question of "what do residents think of these comments", which completed this body of work.

Taken together, each methodologic approach allowed us to see different parts of the phenomenon of how assessment comments are constructed and interpreted. Any one theory, method or approach would have given us only part of the picture, and our understanding is not yet complete even with the studies presented here. Creswell and Plano-Clark's framework for multiphase, mixed-methods research was therefore an ideal fit for this thesis. The last stage, integration, is illustrated best in a separate section after all studies are considered individually, as I've attempted to do throughout this chapter. Integrating the findings under broader themes allowed us to see what cut across the individual studies and what was only visible once all studies were complete: that is, written assessment language is useful, comments are vague but understandable, attendings hedge to save face and that's ok, and subjectivity is not a bad thing.

What did a mixed methods approach obscure or make more difficult?
Despite the numerous advantages to mixed-methods research outlined above and in Chapter 1, using different methods, theories and frameworks can also pose challenges for the researcher. For one thing, it is undoubtedly much more labour intensive to gain sufficient expertise and confidence in a variety of methods. It can also make it more difficult to compare studies directly to one another. As an example, we used a constructivist grounded theory approach to develop a framework to describe how readers decode and interpret written comments. Then we used politeness theory to analyse the comments themselves. Each of these qualitative explorations added to our understanding of why vague language persists, but we couldn't directly compare them, nor could we state with certainty exactly how different politeness strategies were interpreted. For that we would need another study in which we could ask faculty or residents specifically about how they interpreted hedging words, attribution shields and depersonalized language. The use of two frameworks created a more complete picture of the phenomenon of vague language; yet, if we had used a single framework for all studies, we might have been in a better position to compare. Trade-offs such as these are inevitable in any research, and the value of the multiphase, mixed methods approach is that these add-on studies can still be conducted and then integrated into a new understanding. Furthermore, the insights we gained using politeness theory could not have been realized within the grounded theory framework.

Another advantage to a single-method approach might have been the ability to take our framework and see how well it transferred to another setting – for instance, to medical students, or to residents in surgery, family medicine or psychiatry. This might have strengthened generalizability and thus could support more immediate uptake on a practical basis. This research could be done in the future. Still, the advantages to allowing each emerging question to determine the best methodology and approach for the next study seem to outweigh any disadvantages in terms of developing a more holistic approach of language use in written assessment.

Finally, it is worth noting that although the studies presented in this thesis used multiple methods, they focused on one particular assessment instrument: the ITER as used in Internal Medicine at the University of Toronto. We cannot state with certainty that our findings would be transferable to other instruments; however, we can tentatively extrapolate from the literature. ITERs can be considered a type of workplace-based assessment (WBA), of which there are many. Some are more clearly formative in nature, such as the mini-CEX, designed to promote direct observation and allow for immediate feedback. A purely formative written assessment may evoke different patterns of language use than what we found, and may lead to different sorts of interpretations by residents or other faculty. However, the line between formative and summative can become blurred when one uses a programmatic system of assessment. In an assessment programme many instruments and modalities are included, and even if some are meant to be more formative in nature, they all get aggregated, synthesized and integrated to allow for a judgment and decision to be made about a learner's performance. Even presumably low-stakes assessments can be considered high-stakes by learners.[55] This is relevant to our work because written comments on ITERs are, at

least in part, meant to be formative, yet the ITER in our setting also "counts" in a summative way. Thus, although our ITER may be similar enough to other WBA instruments to allow transferability of our findings, these assumptions should be tested in future studies.

Implications and Future Directions

The five studies presented in this thesis suggest that great potential exists for using written comments for assessment purposes. However, this potential is, as yet, untapped. Should we be concerned that educators are not yet fully embracing subjectivity and narrative comments in assessment? After all, the first calls to reconsider subjectivity in assessment came more than 20 years ago.[50,56] Despite the encouraging findings presented in Chapters 2-6, however, it is probably still best to consider narrative descriptions as complementary information to numeric scores. These narratives might serve unique or particular purposes rather than being expected to "replace grades and numerical ratings for clinical performance."[19] We are not, therefore, advocating for a complete overhaul of our assessment systems, although there are intriguing implications that can be considered in the short-term.

There is no doubt that narrative assessment comments show promise, and can be used reliably to distinguish between residents. In Chapter 4 we reported that reliabilities are high enough after three months of comments and with only two raters that it could be a useful method for identifying residents who may need closer attention. And given the findings in Chapter 6, perhaps one of these two raters could be a higher-year resident (e.g., a PGY2 or PGY3 judging PGY1's). This could even have added benefit to the senior residents' own personal and professional development, by learning how assessment is conceptualized and commented on. On the surface it seems that it would be simple enough to institute a process whereby the first three months of ITER comments are collected for each PGY1, put into packages of reasonable numbers (perhaps no more than 16-20), and analyzed and interpreted by faculty or senior residents. However, given the above caveats regarding feasibility, some critical issues should be addressed as programs consider instituting such a process.

One major issue that may limit feasibility is that new resources and infrastructure may be required, to enable collection and collation of these data, to allow for proper anonymization to protect resident and faculty identity, and to develop a system of managing and reporting the outcomes. A residency program would need to determine what such a system would cost, in terms of systems improvements, personnel workload and faculty and residents' time. A determination would then have to be made about whether or not the cost is defensible given a program's size, resources and other demands. It is also not yet clear what a program director should actually do with residents identified as underperforming in some way. Some residents have a slow start but improve with time and experience, while others continue to struggle – these groups may need quite different approaches. Another thing to consider is whether categorizing and rank-ordering, as we did in our studies, is the right approach to take.

In the studies we conducted, this process allowed for excellent discrimination between residents, but that may not be the most important outcome from a resident learning perspective. Further, a forced rank-ordering always results in one resident in last place, but that resident may not actually be underperforming – he or she may be on-track, doing well, in need of some fine tuning, but just not a superstar. Before knowing what attention, resources or interventions might be needed, the program would need to determine its goals and objectives in setting up such a system. If it is meant to flag underperformers early, perhaps a simple categorization system would be more informative and potentially more feasible – such as a green light, yellow light, red light system, anticipating mostly greens, few yellows and rare reds. Importantly, any such analysis and interpretation should look for language indicating exactly what the problem is thought to be, which can then form the basis for any coaching or remediation. In any case, the program and its committee would have to have buy-in and clear goals before instituting such a system, and can look to recent examples in the literature for guidance.[55,57,58]

Given the need for culture-change when it comes to accepting written comments in an assessment program, especially for programs already burdened with impending changes mandated by the shift to competency-based medical education, our findings can add much needed validity evidence to support their use. In modern validity frameworks validity is no longer considered as a property of a test or instrument; rather, educators should think in terms of "validation", a process of collecting and interpreting validity evidence.[59] Validity frameworks are argument-based, meaning that one starts by stating the intended use of a particular instrument and then constructing an argument to support that use.[60] One major advantage of such an approach is that it requires the user to state up-front the purpose of the instrument and what it is to be used for. Then the various stages of the argument can be given relative weight based on that intended use. The framework proposed by Kane is appealing because it can apply equally well to non-numeric data, such as qualitative comments[61] as well as to difficult to assess constructs like professionalism[62] and to programmatic assessment.[63] Our collective findings can be used within this framework to support the validity of using ITER comments for assessment.

The four stages in Kane's validity argument are scoring, generalization, extrapolation and interpretation. Preceding these, though, is identifying the intended use of the instrument. With the ITER in Internal Medicine, this critical first step can actually be problematic, as the ITER is perceived to serve multiple purposes simultaneously.[10,16,48] However, most would agree that the comments are meant to document residents' performance on a particular rotation, and that they "count" for the resident in terms of progression. With that intended use in mind, the first inference is *scoring*, which refers to the process of transforming observations into a score, or in this case into words. Validity evidence for this inference would have to show that observations lead to an insightful and accurate narrative.[59] Our findings support this inference by showing that rich narrative responses are often obtained, relevant issues are focused on, and narratives demonstrate reflexivity of assessors, through their frequent use of hedges and shields. Two other sources of evidence important to scoring are not uni-

versally met – those that support observer credibility and prolonged engagement. These two factors were identified as issues by resident participants in Chapter 6, and our findings are therefore only partially supportive.

The next stage is *generalization*, which refers to how well the "test items" represent all the theoretically possible items in the test universe. As Kane explains, "the observed score is the datum and the universe score is the claim" we make based on that datum.[60] With numeric scores this inference is usually backed by reliability and generalizability studies, and indeed our studies demonstrated high generalizability using both ITER numeric scores and with our rank-order data based on the comments. The *extrapolation* inference looks at how well the generalized interpretations from the test setting extrapolate or map on to real life situations, recognizing that it is not feasible to employ either random or completely representative samples of performance.[60] Our residents' ITER comments do reflect multiple domains relevant to practice, such as knowledge base, communication and work ethic.[20] We also showed in Chapters 4 and 6 that relevant stakeholders, including residents, "agree" with the final interpretations – that is, based only on ITER comments, faculty "outsiders" as well as residents came to the same conclusions about residents as did "internal" faculty. Further, as seen in Chapter 2, both comments and numeric scores were predictive of PGY3 performance. Regarding the final inference, *implications*, our studies do not add relevant evidence as we did not explore ITER comments' role in actual judgments or decisions for individual residents, nor did we focus on potential unintended consequences of such judgments. This inference would be important to generate supporting evidence for, especially if assessment comments are to be used in a summative way.

In summary, taken together, our studies contribute much needed validity evidence that can support the use of written comments as a legitimate component of residents' assessments. Using a validity argument approach to frame our studies also allows educators to note where the arguments and inferences are relatively weaker, which can help researchers prioritize new studies to support the use of comments for summative or other purposes.

Our studies also can have immediate implications for approaches to faculty development. Since success in this regard has been limited despite excellent efforts, it is worth questioning: What is the problem that faculty development is trying to solve? Efforts to date have been focused on getting faculty to write more specific, balanced and behaviourally actionable comments, but as described above, these have had little effect. Not to mention that even if faculty can change the way they write, we don't know what effect that would have on residents or on others who use the comments. Our findings suggest that, in general, the comments in their current state are sufficiently understandable by those who need to read them. And there may be very good reasons why vague language persists and may even be useful. So perhaps new approaches should reinforce the utility that comments have, and the appreciation that residents express for more detail. Further, rather than trying to limit the types of comments commonly deemed to be useless, maybe we should leave those as they are and focus on adding comments that are thought to be helpful. Even here, though,

there are many unanswered questions that would need to be considered first, related to determining what sorts of comments actually lead to better learning.

One area briefly touched on in this thesis is the lack of alignment between ITER-type assessments and competency frameworks. Given what we now know about the reliability and interpretability of written comments, and what that says about a shared understanding of performance, we can use our studies to help improve that alignment. For example, we found that comments often contain language regarding stage of training, such as "already functioning as a consultant", which could reflect supervisors' implicit statements of entrustment. Supervisors often try to convey this information in their comments, as no such language appears in the rating scales. Interestingly, comments related to what level a resident was functioning at were particularly noteworthy as "positive red flags" that readers searched for - and rated highly – when sifting through comments. One possible critique of our work is that we have not explicitly discussed the relationship of written comments to milestones or entrustable professional activities, both important elements of the new competency-based education movement.[64,65] Analyzing the comments for language that reflects entrustment at the various training levels (PGY1-3 in Internal Medicine, for example) can help us determine what "sufficient competence" looks like in order to support promotion to the next level. We can then consider re-aligning our scales to match.[22]

Our studies raise other critical questions that will require further research as we consider incorporating use of narrative comments more fully into assessments. For example, how specific are our findings to internal medicine in Canada? Would similar abilities to interpret comments be seen in other specialties such as surgery, psychiatry or family medicine? Do faculty in other disciplines vary in their use of politeness strategies – are they more or less direct than in Internal Medicine? If so, what effect does that have on how residents' interpret – and learn from – their own comments? This is important in light of the fact that most disciplines have a responsibility to educate and assess trainees from other disciplines. Generating knowledge on discipline-specific differences in assessment language can help residents and educators to make better sense of these cross-discipline assessments. What do we know about the use of assessment language in other countries and cultures? The issue of comment interpretation for IMGs was raised earlier, but what about internationally trained faculty? To what degree does the writing or interpretation of written comments depend in a meaningful way on being a native English language speaker? Similarly, language use and interpretation is known to differ depending on gender, both of the person writing and the person being written about.[66,67] What do we need to understand about gender differences in language use to inform ongoing incorporation of narrative comments into our assessment practices? Indeed, as the research and education communities begin to use comments more explicitly, the data will be generated to allow us to answer many of these questions.

One area that is particularly intriguing, both in terms of immediate implications and for future research, relates to the finding that residents seemed tolerant of staff variability in assessment. One implication is that we can use this finding to support the continu-

ing – and increasing – use of multiple, subjective assessments in our IM residency. Ensuring that residents have ample opportunity to view their comments in aggregate, such as at the 6-month and one-year marks, could allow them to consider what others have written about them in the appropriate context. For example, if a resident had a disappointing comment at the end of one rotation, and dismissed it at the time as being an outlier, viewed in retrospect they may see that there was some merit to that opinion. It would probably be of benefit to have a trusted advisor review the aggregate comments with each resident, to help manage potential emotional reactions[35] as well as to help the resident see the big picture, such as progress over time, or areas still requiring improvement.

The next big question to ask then becomes: Do residents value these variable opinions or merely tolerate them? It is likely that residents' opinions on this would vary. Perhaps those that do value variable opinions would get more benefit from 360° assessments and would be more prepared for lifelong learning. After all, our patients sometimes have variable opinions of us as well. If we can help residents to learn from these "outliers" rather than dismissing them, they may come to see the educational value in these opinions and may find it easier to incorporate others' views as they progress in their careers. In this way, the assessments become more closely aligned to real-life. Knowing more about how (or if) residents value variability in assessment could lend support to the current movements in rater cognition research that emphasize the potential value in "seeing the same thing differently".[52,68] This would then have implications for faculty development. Instead of focusing solely on rater training,[51] we could help faculty better articulate why they've rated a certain way, which could potentially lead to more valuable feedback for their learners.

One final implication of our findings is worth consideration. Elsewhere in this thesis I have discussed the importance of not taking a reductionist approach to language interpretation, instead arguing for the use of whole language in context in order to preserve meaning. But this can raise legitimate concerns about feasibility, especially in large programs, potentially limiting the application of our findings in real life settings. To address the feasibility and workload issues, it is tempting to turn to language software programs that can quickly scan and "score" language on various domains, such as emotional tone, specificity, and dozens of other factors.[69,70] Indeed, the use of such programs was at one point a planned part of this thesis. However, the more I learned, the more I realized there was still much to understand about why and how we use language in our assessment context. Essentially, it was too early to explore these programs because we didn't understand enough about the phenomenon. At this point, however, it may now be worth revisiting such an approach from a research perspective, but informed by what we've learned through our studies. For example, one of the software programs, the Linguistic Index and Word Count,[71] (LIWC) counts the use of pronouns, and we know from Chapter 5 that "depersonalization" (i.e., avoiding the use of names or pronouns) is a prominent politeness strategy. We could explore relationships between depersonalization and how comments are interpreted, and could use the software to assist us in categorizing comments as to their use of pronouns. We also found that the use of "in group identity markers" was significantly more common in

high versus low-rated residents (words such as doctor, consultant, or resident), so their absence might be a signal of lower performance. It is possible that some combination of linguistic features may be able to help us identify residents that require a closer "read", and we can also plan studies to see whether any linguistic features can predict future performance.

Conclusion

In a provocative recent article, Hanson et al. argued that we should abandon numbers and replace them with narratives for clinical performance assessment.[19] Although our research certainly lends support to more fully embracing the use of narratives, I feel this suggestion is premature. Apart from issues of feasibility – and the major culture change that would be required – we do not yet have strong evidence that learning will be improved. Further, our faculty and residents recognized that there are different purposes for the numeric scores and the comments, which supports the continuing use of both in a complementary way. A programmatic view of assessment, using a qualitative research framework to support validity, would allow for the use of multiple sources and types of data when forming judgements and decisions about our trainees.[45,61,72] For this to succeed we would need credible assessors, dedicated time and specific infrastructure and support within the system. We would also need to convince all stakeholders that this is useful, by disseminating and explaining the findings from relevant research and by learning from early-adopter programs. In the meantime, our studies can help encourage faculty in their comment writing, by showing that residents value them, that they would like more specific feedback (either verbal or written) and that comments are potentially more valuable than scores. This seems well within our reach.

References

1. Gray JD. Global rating scales in residency education. *Acad Med*. 1996;71(1):S55-S63.
2. Turnbull J, Gray J, MacFadyen J. Improving in-training evaluation programs. *J Gen Intern Med*. 1998;13(5):317-323.
3. Cohen G, Blumberg P, Ryan N, Sullivan P. Do final grades reflect written qualitative evaluations of student performance? *Teach Learn Med*. 1993;5(1):10-15.
4. Jackson JL, Kay C, Frank M. The validity and reliability of attending evaluations of medicine residents. *SAGE Open Med*. 2015;3:2050312115589648.
5. Canavan C, Holtman MC, Richmond M, Katsufrakis PJ. The quality of written comments on professional behaviors in a developmental multisource feedback program. *Acad Med*. 2010;85(10 Suppl):S106-S109.
6. Ringdahl EN, Delzell JE, Kruse RL. Evaluation of interns by senior residents and faculty: Is there any difference? *Med Educ*. 2004;38(6):646-651.
7. Whitehead CR, Kuper A, Hodges BD, Ellaway R. Conceptual and practical challenges in the assessment of physician competencies. *Med Teach*. 2015;37(3):245-251.
8. Vivekananda-Schmidt P, Mackillop L, Crossley J, Wade W. Do assessor comments on a multi-source feedback instrument provide learner-centred feedback? *Med Educ*. 2013;47(11):1080-1088.
9. Jackson JL, Kay C, Jackson WC, Frank M. The Quality of Written Feedback by Attendings of Internal Medicine Residents. *J Gen Intern Med*. 2015;30(7):973-978.
10. Watling CJ, Kenyon CF, Schulz V, Goldszmidt MA, Zibrowski E, Lingard L. An Exploration of Faculty Perspectives on the In-Training Evaluation of Residents. *Acad Med*. 2010;85(7):1157-1162.
11. Ginsburg S, Regehr G, Lingard L, Eva KW. Reading Between the Lines: Faculty's Interpretations of Narrative Evaluation Comments. *Med Educ*. 2015;49(2):296-306.
12. Patel R, Drover A, Chafe R. Pediatric faculty and residents perspectives on in-training evaluation reports (ITERs). *Can Med Educ J*. 2015;6(2):41-53.
13. Naidich JB, Lee JY, Hansen EC, Smith LG. The Meaning of Excellence. *Acad Radiol*. 2007;14(9):1121-1126.
14. Roberts T. When I say … excellent. *Med Educ*. 2015;49(12):1187-1188.
15. Kiefer CS, Colletti JE, Bellolio MF, et al. The "Good" Dean's Letter. *Acad Med*. 2010;85(11):1705-1708.
16. Ginsburg S, Eva KW, Regehr G. Do in-training evaluation reports deserve their bad reputations? A study of the reliability and predictive ability of ITER scores and narrative comments. *Acad Med*. 2013;88(10):1539-1544.
17. Harris AJL, Corner A, Hahn U. James is polite and punctual (and useless): A Bayesian formalisation of faint praise. *Think Reason*. 2013;19(February 2015):414-429.
18. Ginsburg S, McIlroy J, Oulanova O, Eva KW, Regehr G. Toward authentic clinical evaluation: pitfalls in the pursuit of competency. *Acad Med*. 2010;85(5):780-786.

19. Hanson JL, Rosenberg AA, Lane JL. Narrative descriptions should replace grades and numerical ratings for clinical performance in medical education in the United States. *Front Psychol*. 2013;4:668.
20. Ginsburg S, Gold W, Cavalcanti RB, Kurabi B, McDonald-Blumer H. Competencies "Plus": The nature of written comments on internal medicine residents' evaluation forms. *Acad Med*. 2011;86(10 Suppl):s30-s34.
21. Kennedy TJ, Regehr G, Baker GR, Lingard L. Point-of-care assessment of medical trainee competence for independent clinical work. *Acad Med*. 2008;83(10 Suppl):S89-S92.
22. Crossley J, Johnson G, Booth J, Wade W. Good questions, good answers: construct alignment improves the performance of workplace-based assessment scales. *Med Educ*. 2011;45(6):560-569.
23. Zibrowski EM, Singh SI, Goldszmidt MA, et al. The sum of the parts detracts from the intended whole: competencies and in-training assessments. *Med Educ*. 2009;43(8):741-748.
24. Crossley J, Jolly BC. Making sense of work-based assessment: ask the right questions, in the right way, about the right things, of the right people. *Med Educ*. 2012;46(1):28-37.
25. Scarcella R, Brunak J. On speaking politely in a second language. *Int J Soc Lang*. 1981;27(1):59-76.
26. Danesi M. Metaphorical competence in second language acquisition and second language teaching: The neglected dimension. In: Alatis JE, ed. *Georgetown University Round Table on Languages and Linguistics 1992: Language, Communication and Social Meaning*. Washington, DC: Georgetown University Press; 1993:489-500.
27. Dahm MR, Yates L, Ogden K, Rooney K, Sheldon B. Enhancing international medical graduates' communication: the contribution of applied linguistics. *Med Educ*. 2015;49(8):828-837.
28. Schabort I, Mercuri M, Grierson L. Predicting international medical graduate success on college certification examinations: responding to the Thomson and Cohl judicial report on IMG selection. *Can Fam physician Médecin Fam Can*. 2014;60(10):e478-e484.
29. Brown P, Levinson SC. *Politeness : Some Universals in Language Usage*. (Levinson SC, ed.). New York: Cambridge University Press; 1987.
30. Ilott I, Murphy R. Feelings and failing in professional training: The assessor's dilemma. *Assess Eval High Educ*. 1997;22(3):307-316.
31. Dudek NL, Marks MB, Regehr G. Failure to fail: the perspectives of clinical supervisors. *Acad Med*. 2005;80(10 Suppl):S84-S87.
32. Ginsburg S, van der Vleuten CPM, Lingard L. Hedging to save face: A linguistic analysis of ITER comments. *Adv Heal Sci Educ*. 2015;Online Ear.
33. Stroud L, Bryden P, Kurabi B, Ginsburg S. Putting performance in context: the perceived influence of environmental factors on work-based performance. *Perspect Med Educ*. 2015;4(5):233-243.
34. Parkes J, Abercrombie S, McCarty T. Feedback sandwiches affect perceptions but not performance. *Adv Heal Sci Educ*. 2013;18(3):397-407.
35. Sargeant J, Mann K V, Sinclair D, van der Vleuten CPM, Metsemakers J.

Understanding the influence of emotions and reflection upon multi-source feedback acceptance and use. *Adv Heal Sci Educ*. 2008;13(3):275-288.
36. Lefroy J, Watling C, Teunissen PW, Brand P. Guidelines: the do's, don'ts and don't knows of feedback for clinical education. *Perspect Med Educ*. 2015;4:284-299.
37. Boehler ML, Rogers DA, Schwind CJ, et al. An investigation of medical student reactions to feedback: a randomised controlled trial. *Med Educ*. 2006;40(8):746-749.
38. Dudek NL, Marks MB, Wood TJ, Lee AC. Assessing the quality of supervisors' completed clinical evaluation reports. *Med Educ*. 2008;42(8):816-822.
39. Dudek NL, Marks MB, Wood TJ, et al. Quality evaluation reports: Can a faculty development program make a difference? *Med Teach*. 2012;34(11):e725-e731.
40. Watling CJ. Cognition, culture, and credibility: deconstructing feedback in medical education. *Perspect Med Educ*. 2014;3(2):124-128.
41. Eva KW, Armson H, Holmboe E, et al. Factors influencing responsiveness to feedback: on the interplay between fear, confidence, and reasoning processes. *Adv Heal Sci Educ*. 2012;17(1):15-26.
42. Eva KW, Bordage G, Campbell C, et al. Towards a program of assessment for health professionals: from training into practice. *Adv Heal Sci Educ*. 2015;Online ear.
43. Schuwirth LW, Colliver J, Gruppen LD, et al. Criteria for good assessment: Consensus statement and recommendations from the Ottawa 2010 conference. *Med Teach*. 2011;33(3):206-214.
44. Hodges BD. Assessment in the post-psychometric era: Learning to love the subjective and collective. *Med Teach*. 2013;35(7):564-568.
45. Govaerts MJB, van der Vleuten CPM, Schuwirth LW, Muijtjens AM. Broadening perspectives on clinical performance assessment: rethinking the nature of in-training assessment. *Adv Heal Sci Educ*. 2007;12(2):239-260.
46. Eva KW, Hodges BD. Scylla or Charybdis? Can we navigate between objectification and judgement in assessment? *Med Educ*. 2012;46(9):914-919.
47. Littlefield JH, Darosa D, Paukert J, Williams RG, Klamen DL, Schoolfield JD. Improving resident performance assessment data: numeric precision and narrative specificity. *Acad Med*. 2005;80(5):489-495.
48. Watling CJ, Kenyon CF, Zibrowski EM, et al. Rules of engagement: residents' perceptions of the in-training evaluation process. *Acad Med*. 2008;83(10 Suppl):S97-S100.
49. Williams RG, Klamen DL, McGaghie WC. Cognitive, Social and Environmental Sources of Bias in Clinical Performance Ratings. *Teach Learn Med*. 2003;15(4):270-292.
50. van der Vleuten CPM, Norman GR, Graaff E. Pitfalls in the pursuit of objectivity: issues of reliability. *Med Educ*. 1991;25(2):110-118.
51. Kogan JR, Conforti LN, Bernabeo E, Iobst W, Holmboe E. How faculty members experience workplace-based assessment rater training: a qualitative study. *Med Educ*. 2015;49(7):692-708.
52. Gingerich A, Kogan JR, Yeates P, Govaerts MJB, Holmboe E. Seeing the "black box" differently: assessor cognition from three research perspectives. *Med*

Educ. 2014;48(11):1055-1068.
53. van der Vleuten CPM, Schuwirth LW, Driessen EW, et al. A model for programmatic assessment fit for purpose. *Med Teach*. 2012;34(3):205-214.
54. Creswell JW, Klassen AC, Plano VL, Smith KC. *Best Practices for Mixed Methods Research in the Health Sciences. A Report Commissioned by the Office of Behavioural and Social Sciences Research*.; 2011.
55. Heeneman S, Oudkerk Pool A, Schuwirth LW, van der Vleuten CPM, Driessen EW. The impact of programmatic assessment on student learning: theory versus practice. *Med Educ*. 2015;49(5):487-498.
56. Norman GR, van der Vleuten CPM, Graaff E. Pitfalls in the pursuit of objectivity: issues of validity, efficiency and acceptability. *Med Educ*. 1991;25(2):119-126.
57. Driessen EW, Tartwijk J van, Govaerts M, Teunissen P, Vleuten CPM van der. The use of programmatic assessment in the clinical workplace: A Maastricht case report. *Med Teach*. 2012;34(3):226-231.
58. Bok HGJ, Teunissen PW, Favier RP, et al. Programmatic assessment of competency-based workplace learning: when theory meets practice. *BMC Med Educ*. 2013;13(1):123.
59. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: a practical guide to Kane's framework. *Med Educ*. 2015;49(6):560-575.
60. Kane M. The argument-based approach to validation. *School Psych Rev*. 2013;42(4):448-457.
61. Cook DA, Kuper A, Hatala R, Ginsburg S. When assessment data are words: Validity evidence for qualitative educational assessments. *Acad Med*. 2016;(In press).
62. Mazor KM, Canavan C, Farrell M, Margolis MJ, Clauser BE. Collecting validity evidence for an assessment of professionalism: findings from think-aloud interviews. *Acad Med*. 2008;83(10 Suppl):S9-S12.
63. Schuwirth LW, van der Vleuten CPM. Programmatic assessment and Kane's validity perspective. *Med Educ*. 2012;46(1):38-48.
64. Ten Cate OT, Scheele F. Competency-based postgraduate training: can we bridge the gap between theory and clinical practice? *Acad Med*. 2007;82(6):542-547.
65. Hawkins RE, Welcher CM, Holmboe ES, et al. Implementation of competency-based medical education: are we addressing the concerns and challenges? *Med Educ*. 2015;49(11):1086-1102.
66. Axelson RD, Solow CM, Ferguson KJ, Cohen MB. Assessing implicit gender bias in Medical Student Performance Evaluations. *Eval Health Prof*. 2010;33(3):365-385.
67. Kaatz A, Magua W, Zimmerman DR. A Quantitative Linguistic Analysis of National Institutes of Health R01 Application Critiques From Investigators at One Institution. *Acad Med*. 2015;90(1):69-75.
68. Yeates P, O'Neill P, Mann K V, Eva KW. Seeing the same thing differently : Mechanisms that contribute to assessor differences in directly-observed performance assessments. *Adv Heal Sci Educ*. 2013;18(3):325-341.
69. Tausczik YR, Pennebaker JW. The Psychological Meaning of Words: LIWC and

Computerized Text Analysis Methods. *J Lang Soc Psychol*. 2010;29(1):24-54.
70. Whissell C. Using the Revised Dictionary of Affect in Language to quantify the emotional undertones of samples of natural language . *Psychol Rep*. 2009;105(2):509-521.
71. Pennebaker JW, Chung CK, Ireland M, Gonzales A, Booth RJ. The Development and Psychometric Properties of LIWC 2007.
72. Driessen EW, van der Vleuten CPM, Schuwirth LW, van Tartwijk J, Vermunt JD. The use of qualitative research criteria for portfolio assessment as an alternative to reliability evaluation: a case study. *Med Educ*. 2005;39(2):214-220.

Summary

Subjectivity in assessment is gaining increasing respect in the medical education community. The overall goal of this proposed research program is to view subjectivity through the lens of language – what we say and how we say it can provide a window through which we can start to see how clinical supervisors construct opinions and judgments about their learners. Analyzing the language assessors use can deepen our understanding of how they conceptualize competence and performance. Learning how others interpret that language can provide necessary evidence to support the validity of using narrative comments in a way that is credible and defensible.

The **Introduction chapter** sets the stage for the five studies that follow by critically reviewing the literature on the use of written comments in the assessment of learners in health professions education. These written comments can serve as a lens with which to better understand how clinical supervisors subjectively conceptualize residents' performance. Several approaches to analyzing assessment language are described, along with a consideration of what these approaches might offer to our understanding. The overall goals for this thesis were twofold: to determine if narrative comments could be used for assessment in a way that is reliable, credible and has validity for its intended purpose; and to gain a deeper and more nuanced understanding of the language attendings use when assessing their residents in Internal Medicine. Given the educational potential of written assessment comments it is important to explore how comments are constructed, why attendings write the way they do, what their language means to others and what that might say about assessment as a whole. Together, the studies in this thesis form a multiphase, mixed-methods program of research.

The study reported in **Chapter 2** used a database of written comments on Internal Medicine residents' in-training evaluation reports (ITERs) from one program in Canada. We found that it was possible for faculty participants to discriminate between these residents based only on the comments, with excellent inter-rater reliability. Both written comments and numeric scores on the ITER in the first postgraduate year (PGY1) were predictive of performance in PGY3 and comments had fairly high correlations with assigned scores within each year. However, because the ITER scores were already fairly predictive of PGY3 performance the comments did not add additional value when both comments and scores were included in a regression model. Still, this study did show that resident performance can be captured through narrative comments in a reliable way.

To understand *how* faculty were able to rank-order comments so reliably – despite the sometimes vague nature of the language they contain – we conducted the study reported in **Chapter 3**, which involved a constructivist grounded theory analysis of interviews with 24 faculty participants. We used constructivist grounded theory in this study because our goal was to develop a framework to explain the process of interpretation. Faculty did not interpret language at face value; rather, they read between the lines to decode the language in an active process to construct meaning from the comments. Their ability to interpret the comments so reliably suggests a shared under-

standing of a "hidden code" in the language used to describe resident performance. This understanding was not perfect, however, and faculty did express difficulties in interpreting vague, generic and dispositional language. This study raised several critical questions which we then attempted to address in subsequent studies, as follows: (i) To what extent is the code universal versus locally specific, for example to a particular residency program or institution? (Chapter 4); (ii) How much written commentary is required to give a stable impression of a resident using this decoding mechanism? (Chapter 4); (iii) Why does such a "hidden code" exist and what purpose might it be serving? (Chapter 5); (iv) Do residents also know the code and can they interpret the comments effectively and reliably? (Chapter 6).

In **Chapter 4** we addressed the first two questions raised at the end of Chapter 3. To do this we enrolled 24 "outsiders", that is, faculty participants from academic departments of medicine across Canada external to our institution. This time we used as our dataset two cohorts of PGY1 residents' comments; faculty were asked to rank-order a set of residents based on either an entire years' worth of comments or based on only the first three months. We found that it did not matter if faculty were insiders or outsiders – reliabilities remained high, suggesting a degree of universality to the "hidden code" in the comments. Further, using only the first three months of comments yielded comparable reliabilities to using the entire years' worth. Decision studies suggest that acceptable reliability can be achieved using two faculty raters and only the first three months of assessment comments.

In **Chapter 5** we explored the question of *why* attendings write the way they do, with a preponderance of vague, generic and dispositional comments. For this purpose we turned to linguistic pragmatics, which focuses on how people use and understand non-literal language. In particular, we used politeness theory, which posits that people use strategies in their communication in order to allow others (or themselves) to "save face". The pervasive use of "hedging" language suggested to us that writing ITER comments is a face-threatening act, for a number of potential reasons outlined in the chapter. Linguists view politeness and hedging as crucial to enabling smooth social interactions and do not consider these strategies in themselves to be problematic. This may explain, in part, why efforts to prompt faculty to write more balanced and critical comments have been met with little success – this social lubrication is necessary to allow faculty attendings to fulfill their various roles: as teachers, mentors, colleagues and assessors. Faculty's use of politeness strategies may also reflect the culture in which we operate and conformity to the norms that exist in the education context. Yet, although politeness in itself is not necessarily problematic, non-literal language use may lead to misinterpretation – and do not yet know how residents interpret their assessment comments.

This last question, also raised in Chapter 3, is addressed in the final study, reported in **Chapter 6**, which completes the story by exploring residents' understanding of ITER comments. For this purpose we recruited 12 PGY2s from our own IM program and replicated the protocol described in Chapter 2. We found that these residents were able to discriminate between PGY1s based on comments alone with extremely high

inter-rater reliability and that the correlation with faculty rankings of the same residents was nearly perfect. Similar to faculty, residents did not take language at face value but made interpretations and inferences, in much the same way that faculty did. They did not seem to misinterpret the messages despite the pervasive use of non-literal language, which should be reassuring to educators. One unexpected finding was that while residents acknowledged and commented on variability between attendings, they seemed to treat it rather nonchalantly, which may lend some support to the continuing push to embrace more subjectivity in assessment.

The **Discussion chapter** comprises the integration stage of the multiphase, mixed-methods program of research presented in Chapters 2-6. During the process of integration and synthesis the findings from all five studies were considered together and four major themes were identified. First, written assessment comments are useful and should no longer be neglected as valuable sources of data. A re-appraisal of existing research in light of our findings helps explain apparent discrepancies between other researchers' findings and our own. In particular, we problematize the concepts of language specificity and feedback in the setting of assessment. Second, assessment language can be vague but is still decodable – the "hidden code" is accessible. The advantages of studying language in context are also discussed. Third, there is a powerful need to "save face" in assessments, which should not be underestimated. Understanding the social value of politeness can lead to new faculty development approaches. Finally, our findings can lend support to the ongoing push towards embracing subjectivity and collectivity in assessment. This chapter also considers the strengths and limitations of using a mixed-methods approach within a program of research. Finally, implications for practice and future directions for research are presented, including a consideration of the validity evidence supporting the use of written comments in resident assessment.

Samenvatting

In de medische onderwijsgemeenschap wordt steeds vaker aandacht besteed aan subjectiviteit bij toetsing. In brede zin beoogt het aan u voorgelegde onderzoeksprogramma subjectiviteit vanuit een taaloogpunt te benaderen; wat we zeggen en de manier waarop we dat doen kan ons aanknopingspunten aanreiken voor een beter begrip van de manier waarop klinisch begeleiders meningen en oordelen vormen over hun studenten. Door het taalgebruik van beoordelaars te analyseren, wordt mogelijk duidelijker hoe zij zich een beeld vormen van bekwaamheid en functioneren. Meer inzicht in hoe anderen die taal interpreteren kan ons het bewijs leveren dat nodig is om de validiteit ten aanzien van het gebruik van narratieve feedback zodanig te onderschragen dat het geloofwaardig en verdedigbaar is.

Het **inleidend hoofdstuk** bereidt de lezer voor op de vijf daaropvolgende studies middels een kritische uiteenzetting van de literatuur over het gebruik van geschreven feedback bij het beoordelen van studenten in het gezondheidszorgonderwijs. Deze geschreven feedback kan als springplank fungeren naar een beter begrip van de manier waarop begeleiders zich op subjectieve wijze een beeld vormen van het functioneren van artsen in opleiding (aiossen). Er worden verschillende benaderingen beschreven voor het analyseren van "toetstaal", welke vervolgens naar waarde worden geschat in een beschouwing van hun eventuele bijdrage aan ons begrip. Het algemene doel van dit proefschrift was tweeledig: 1) te bepalen of narratieve feedback op een betrouwbare, geloofwaardige en valide manier kan worden aangewend voor toetsdoeleinden, en 2) een beter en genuanceerder begrip te krijgen van het taalgebruik van toeziend artsen bij het beoordelen van artsen in opleiding tot internist. Omdat geschreven feedback op het functioneren een positieve rol kan spelen in het leerproces is het belangrijk te onderzoeken hoe die feedback wordt opgebouwd, waarom toeziend artsen schrijven zoals ze schrijven, wat hun taal betekent voor anderen en welke gevolgtrekkingen ten aanzien van toetsing we daaruit zouden kunnen maken. De studies in dit proefschrift vormen samen een uit diverse fasen bestaand (*multiphase)* onderzoeksprogramma waarbij zowel kwalitatieve als kwantitatieve methoden zijn toegepast (*mixed methods*).

In de in **Hoofdstuk 2** gerapporteerde studie werd gebruik gemaakt van geschreven feedback op de evaluatieverslagen van artsen in opleiding tot internist (ITERs*) van één opleiding in Canada. Onze bevinding was dat deelnemende stafleden in staat waren onderscheid te maken tussen deze aiossen op basis van feedback alleen en daarbij een uitstekende interbeoordelaarsbetrouwbaarheid vertoonden. Zowel de geschreven feedback op de ITERs in het eerste jaar van de vervolgopleiding als de daaraan toegekende numerieke scores bleken voorspellers te zijn van het functioneren in het 3e studiejaar. Ook bleek dat binnen elk jaar de feedback vrij sterk correleerde met de toegekende scores. Omdat de ITER-scores echter al een redelijke voorspellende waarde hadden voor het functioneren in het 3e studiejaar, bleek nadat we feedback en

scores hadden opgenomen in een regressiemodel dat de feedback geen extra waarde toevoegde. Desalniettemin werd met deze studie aangetoond dat het functioneren van aiossen op betrouwbare wijze in beeld kan worden gebracht met behulp van narratieve feedback.

Om erachter te komen hoe het kón dat stafleden de feedback op zo een betrouwbare wijze wisten te rangschikken, ondanks de vage taal die deze soms bevatte, verrichtten we de in **Hoofdstuk 3** vermelde studie waarbij we aan de hand van een constructivistische gefundeerde theoriebenadering interviews met 24 deelnemende stafleden analyseerden. We kozen voor een constructivistische gefundeerde theoriebenadering, omdat we ten doel hadden een kader te ontwikkelen waarmee het interpretatieproces kon worden verklaard. Stafleden bleken de taal niet letterlijk te interpreteren; in plaats daarvan lazen ze tussen de regels door om de taal middels een actief proces te kunnen ontcijferen en betekenis te kunnen abstraheren uit de feedback.

Het feit dat zij zo goed in staat waren de feedback op betrouwbare wijze te interpreteren maakt het aannemelijk dat zij allen een "verborgen taal" beheersten waarmee zij specifiek het functioneren van aiossen beschreven. Deze taalbeheersing was echter niet perfect, daar stafleden te kennen gaven hier en daar moeite te hebben met het interpreteren van vaag, algemeen en persoonsgebonden taalgebruik. De studie riep enkele belangrijke vragen op die we in de volgende studies trachtten te beantwoorden, namelijk: (i) In hoeverre is deze verborgen taal universeel dan wel plaatsgebonden, bijvoorbeeld aan een bepaalde vervolgopleiding of instelling? (Hoofdstuk 4); (ii) Hoeveel geschreven feedback is er nodig om een stabiel beeld te geven van een aios middels dit ontcijfermechanisme? (Hoofdstuk 4); (iii) Waarom bestaat er zo een "verborgen taal" en wat zou het doel ervan kunnen zijn? (Hoofdstuk 5); (iv) Kennen aiossen deze taal ook en weten zij de feedback op efficiënte en betrouwbare wijze te interpreteren? (Hoofdstuk 6).

In **Hoofdstuk 4** werden de eerste twee vragen beantwoord die aan het eind van Hoofdstuk 3 werden gesteld. Daartoe nodigden we 24 "buitenstaanders" uit, dus deelnemende stafleden van academische geneeskundeafdelingen uit heel Canada die niet aan onze instelling verbonden waren. Onze dataset bestond ditmaal uit de feedback van twee cohorten eerstejaars aiossen; we vroegen stafleden een aantal aiossen te rangschikken op basis van alle gedurende een jaar ontvangen feedback of op basis van feedback over alleen de eerste drie maanden. Onze bevinding was dat het niet uitmaakte of stafleden nu binnen- of buitenstaanders waren: de betrouwbaarheid bleef hoog, wat erop duidde dat de "verborgen taal" in de feedback in zekere mate universeel was. Verder maakte het voor de betrouwbaarheid vrijwel geen verschil of de beoordeling berustte op feedback over alleen de eerste drie maanden of op feedback over het gehele jaar. Studies over besluitvorming tonen aan dat een acceptabele be-

trouwbaarheid bereikt kan worden wanneer er twee stafleden worden ingezet als beoordelaar en wanneer de feedback op het functioneren alleen de eerste drie maanden betreft.

In **Hoofdstuk 5** onderzochten we de vraag waarom toeziend artsen schrijven zoals ze schrijven met overwegend vaag, algemeen en persoonsgebonden feedback. Wij zochten het antwoord in de linguïstische pragmatiek die zich richt op hoe mensen niet-letterlijke taal gebruiken en begrijpen. We maakten voornamelijk gebruik van beleefdheidstheorie waarbij gesteld wordt dat mensen strategisch communiceren om te voorkomen dat anderen (of zijzelf) "gezichtsverlies" lijden. Het veelvuldige gebruik van "indekkende" taal deed ons veronderstellen dat het schrijven van ITER-feedback een gezichtsbedreigende aangelegenheid is, om een aantal mogelijke redenen die in het hoofdstuk worden uiteengezet. Taalkundigen beschouwen beleefdheid en indekken als essentieel voor een soepel verloop van sociaal contact en zien geen kwaad in het gebruik van communicatiestrategieën op zichzelf. Dit zou (deels) kunnen verklaren waarom pogingen om stafleden ertoe aan te zetten meer gebalanceerde en kritische feedback te schrijven tot nu toe weinig succes hebben gehad: toeziend artsen hebben deze sociale smering nodig om hun verschillende rollen als docent, mentor, collega en beoordelaar te kunnen vervullen. Het gebruik van beleefdheidsstrategieën door stafleden kan ook een weerspiegeling zijn van de cultuur waarin we ons begeven en van naleving van de normen die in de onderwijscontext gelden. Toch moeten we er rekening mee houden dat, ook al hoeft beleefdheid op zich niet problematisch te zijn, niet-letterlijk taalgebruik verkeerde interpretaties in de hand kan werken. Daarnaast weten we nog niet hoe aiossen zelf de feedback op hun functioneren interpreteren.

Deze laatste vraag die ook in Hoofdstuk 3 werd gesteld, wordt beantwoord in de laatste, in **Hoofdstuk 6** beschreven studie, welke het verhaal afsluit met een nader onderzoek van het begrip dat aiossen hebben van ITER- feedback. Voor dit doel nodigden we 12 tweedejaars aiossen van onze eigen interne geneeskundeopleiding uit en voerden we dezelfde stappen uit als die we in Hoofdstuk 2 beschreven. Onze bevinding was dat deze aiossen in staat waren onderscheid te maken tussen eerstejaars aiossen op basis van alleen feedback en daarbij een buitengewoon hoge interbeoordelaarsbetrouwbaarheid vertoonden. Bovendien was de correlatie met de rangorde die stafleden gemaakt hadden van dezelfde aiossen vrijwel perfect. Net als de stafleden namen de aiossen de feedback niet letterlijk, maar interpreteerden deze en trokken conclusies op nagenoeg dezelfde wijze als stafleden dat deden. Niets wees erop dat zij het commentaar verkeerd interpreteerden, ondanks het veelvuldige gebruik van niet-letterlijke taal, en dit zou opleiders gerust moeten stellen. Wat we niet verwachtten te vinden was dat aiossen nogal onverschillig stonden ten opzichte van de verschillen tussen toeziend artsen, terwijl zij wel degelijk erkenden dat ze bestonden en dit ook opmerk-

ten. Deze bevinding kan de aanhoudende vraag naar meer subjectiviteit bij toetsing kracht bijzetten.

Het **Discussiehoofdstuk** beslaat de laatste fase van het *multiphase*, *mixed-methods* onderzoeksprogramma tijdens welke de in Hoofdstukken 2 t/m 6 gepresenteerde fasen worden geïntegreerd. Tijdens het integratie- en syntheseproces werden de bevindingen van alle vijf de studies in hun geheel beschouwd en daarbij kwamen vier hoofdthema's naar voren. Ten eerste is geschreven feedback op het functioneren bruikbaar en deze zou als waardevolle gegevensbron niet meer onbeschouwd mogen blijven. Door bestaand onderzoek opnieuw te bekijken in het licht van onze bevindingen, kunnen duidelijke afwijkingen tussen de bevindingen van andere onderzoekers en de onze worden verklaard. Zo maken we bijvoorbeeld teveel een probleem van concepten als taalspecificiteit en feedback als het gaat om toetsing. Ten tweede mag toetstaal dan wel vaag zijn, maar het is toch te ontcijferen; met andere woorden: men heeft toegang tot de "verborgen taal". De voordelen van het bestuderen van taal in de context worden ook besproken. In de derde plaats bestaat er een sterke behoefte aan het voorkomen van "gezichtsverlies" bij toetsing, welke niet onderschat moet worden. Een beter begrip van de sociale waarde van beleefdheid kan ons helpen docentprofessionaliseringsprogramma's te vernieuwen. Ten slotte kunnen onze bevindingen het draagvlak voor subjectiviteit en collectiviteit bij toetsing verbreden. Dit hoofdstuk bespreekt ook de voor- en nadelen van een *mixed methods*-benadering binnen een onderzoeksprogramma. Als laatste worden de gevolgen voor de praktijk en aanbevelingen voor toekomstig onderzoek gepresenteerd, samen met een beschouwing van het bewijs dat de validiteit van toetsing van aiossen op basis van geschreven feedback bij de beoordeling van aiossen aantoont en het gebruik daarvan ondersteunt.

Valorization

1. **(Relevance) What is the social (and/or economic) relevance of your research results (i.e. in addition to the scientific relevance)?**

The results from this research can have potential economic advantages for medical schools and residency training programs. As described in Chapters 2 and 4 we found that it is possible to rank-order internal medicine residents (from highest to lowest) based on an analysis of their supervisors' assessment comments and that this rank-ordering is extremely reliable. Further, we determined in Chapter 4 that this high reliability can be shown even after only 3 months of residents' clinical rotations, with 2 faculty members analyzing and rank-ordering – far exceeding the results based on numeric scores alone.
Residents ranked as lowest in their cohort(s) can therefore be identified much earlier using this method when compared to looking only at their numeric scores. The low ranked residents can then be flagged for closer attention, follow-up or even remediation if necessary. This would not be possible using the numeric scores until much later – perhaps at 6 months or even the end of the year. Thus there is the potential – with relatively modest time and effort – to act early and get residents on track for success. If remediation is needed it can begin immediately and this can save time and effort at the end of the program.

Furthermore, since we also found (in Chapter 6) that senior residents in the same Internal Medicine could analyze and rank-order written comments as reliably as faculty could, a senior resident could be paired with a faculty member for the task described above, which would further decrease the time and costs involved. It would also have the added benefit of engaging senior residents in an educational process, which could allow them to reflect on their own assessments and which also might spark an interest in educational development or scholarship.

Another benefit of being able to rigorously use written comments for assessment is that it helps avoid the reductionist approach of numeric scoring. When using only numeric scoring and checklists the clinician-supervisor must observe a performance and then mentally translate that into one or more of several possible competencies (e.g., did the performance relate to physical exam skills, communication or professionalism?) following which they must choose a score from 1-5 to represent their assessment. More nuanced and subtle information can get lost in these multiple translations, which affects the authenticity and validity of the assessment process. On the other hand, if the performance can be captured in words, using natural language, it is likely that more "raw" information or data will be preserved. This assessment language can be used to create more robust documentation (beyond numeric scores), which can be critically important to support (or defend) a learner's remediation or dismissal from a program. This is a notoriously difficult, time consuming and expensive process, and any developments to reduce unnecessary time and costs will be of great value.

2. (Target groups) To whom, in addition to the academic community, are your research results of interest and why?

One important target audience is the promotions committees at medical schools and residency programs. These committees are becoming more prevalent as medical education shifts towards a competency-based (rather than strictly time-based) approach. Promotion committees assess a learner's progress and trajectory and decide whether or not they are ready to move to the next stage of training. Since much of the data collected is in the form of words (assessment comments) our research would be of great interest in that it can lead to a more systematic and rigorous method for analyzing these comments. Our findings also provide evidence to support the validity of such an approach, which is critical in ensuring buy-in and in case evidence is needed for an appeal.

The approach we developed – and the new knowledge generated regarding the value of written comments – will also be very helpful to organizations such as the Royal College of Physicians and Surgeons of Canada. As mentioned, with the shift to CBME there will be many more assessments gathered on each resident, much of which may be in the form of written comments. The Royal College also still promotes the use of In-Training Evaluation Reports (ITERs) which were the subject of all of the studies comprising this thesis. Our findings therefore have great potential to inform and guide the RC in determining best practices for writing and analyzing written comments.

Finally, the resident learners themselves are an important target audience. From residents' perspectives, we found that assessment comments may be preferable to numeric scores, or at the least they are considered necessary and complementary (see Chapter 6). In multiple studies researchers have found that resident learners often express frustration about their assessments, reporting that they are not useful to their development. Our findings suggest that if the assessments include more and "better" comments that residents will find them more authentic and helpful. This may lead to better engagement in their assessment process and more "buy-in" with regards to their own feedback. Further research would have to be done, but this system of assessment could have effects on residents' self-directed learning and self-assessment abilities, which are very important to ongoing maintenance of competence.

3. (Activities/Products) Into which concrete products, services, processes, activities or commercial activities will your results be translated and shaped?

Our findings, along with the suggestions we recommend in the Discussion chapter, can be immediately implemented and tested for validity and feasibility. In Internal Medicine, at least in Canada, the writing and interpretation of assessment comments about residents seems fairly universal; that is, internal medicine attending physicians from

across the country came to the same judgments about residents based on reading their assessment comments. We believe that at least in Internal Medicine in Canada it will be possible to transfer these findings to programs across the country, who can begin to use assessment comments in a systematic way. It is important to note, however, that the process and model for assessing comments may not be immediately or directly transferable to other programs, such as surgery or psychiatry, or in other settings or countries, and would have to be rigorously tested prior to use. It would have to be shown that comments from these domains can be interpreted reliably and with similar predictive value when compared to our findings in Internal Medicine. Any program, however, can take our published protocol and easily adapt it for their own research.

Another demonstration of the translation of this research is in its dissemination. For example, Chapters 2, 3 and 5 have already been published in the peer-reviewed literature (Chapter 2 is also available free online from the journal) and have received approximately 20 citations so far. Chapter 5 just came out in print yet has been read nearly 100 times so far on researchgate.net, indicating that it is gathering significant interest. I have already spoken publicly about my research at other universities in Canada and have more public speaking engagements scheduled in the next few months, all of which will serve to disseminate and promote our findings more broadly.
Chapters 4 and 6 are currently under editorial review at two different journals.

A further indication of the potential translation of this work can be found by considering the interest this research has garnered at the organizational level. For example, I have been invited to sit on a newly formed research advisory committee at the Medical Council of Canada, the organization that develops and administers the examinations required prior to obtaining a medical license to practice in Canada. This new committee is tasked with advising the MCC, it's psychometric and research committees about ongoing and future assessment processes necessary to enable the MCC to meet its important goals. My participation will provide an important venue for translating our research findings into practice and also in developing and conducting new studies to further refine and test our findings in multiple programs.

4. (Innovation) To what degree can your results be called innovative in respect to the existing range of products, services, processes, activities and commercial activities?

Our research is innovative because it is the first to show, in a systematic and rigorous way, that written assessment comments about medical trainees can be highly reliable and useful for assessment. Our findings bolster the validity argument supporting the use of such assessments for determining the competence of our trainees. Prior to this, nearly all of our assessments in medical education have been based on numeric scores, which have their strengths and important uses but also have many limitations. This

research therefore advances the field in a way that could have broad implications for any other program that uses workplace based assessment.

5. (Schedule & Implementation) How will this/these plan(s) for valorization be shaped? What is the schedule, are there risks involved, what market opportunities are there and what are the costs involved?

Dissemination of the novel research in this thesis has already begun. In addition to the publications noted above, the thesis book will also be published and publicly available in the fall of 2016. I have begun speaking about this research at local and national conferences and events. In terms of uptake, it will of course take time for these new ideas to be adopted by others but we hope that our research will be replicated in other disciplines. When I present this work publicly I always invite interested researchers or stakeholders to contact me if they are interested in our protocol and I have offered assistance in setting up similar studies elsewhere. As it is quite new it is not yet clear how long it may take until replication studies can be completed, but there is great potential in the next one to two years for this to occur.

At this point it is difficult to foresee immediate market opportunities related to this research and thus it is not possible to estimate costs. However, one relative risk is worth noting as we consider shaping valorization plans and that relates to the prevailing quantitative culture at most institutions. As mentioned in this section and throughout this dissertation (see especially the Discussion section) our medical education systems have been based almost exclusively on numeric scoring systems for assessment. The administrative and regulatory systems rely heavily on gathering and reporting numbers and scores, and are not optimally structured to allow for the collection and interpretation of qualitative data. For new systems – based on qualitative assessment – to not only be implemented but to succeed and reach their full potential will require a significant shift in mindset and culture. This can take significant effort and time but we are confident that this effort will ultimately be fruitful.

Acknowledgments

Many people have asked what possessed me to enrol in a PhD program at this stage of my life. That is a question with many answers! But I wouldn't have even considered it if not for the incredible support and patience of my family. My husband Stuart was not the least bit surprised when I first hinted that I wanted to do a PhD. He always knew it was in me somewhere and I am forever grateful for his partnership, love and support. My two daughters, Maya and Sage, were my own personal cheering squad and were always proud of me, no matter how obsessed I became with my writing at times. Which I did.

I truly had a dream team of supervisors guiding me through the last few years. They are such giants in the medical education research community and I was thrilled when they agreed to take me on as a student. I have learned so much from each of them. No one can match Cees van der Vleuten's turnaround time for feedback, which was all the more remarkable for the geographic diversity from which it arose. It was always enthusiastic while appropriately critical, and greatly shaped my way of thinking. I have incredible admiration for Kevin Eva's patience as I tried to learn quantitative methods and as I learned to embrace the role of student. His critical eye taught me to write with precision and to think like an editor. Lorelei Lingard, our field's most stylish writer in every sense, was phenomenally gracious and generous with her feedback to me as I (hopefully) learned about the power of verbs, the proper use of commas and the importance of parallel structure. Her love of language is infectious and was a great source of inspiration for my work.

A special thanks also goes to Glenn Regehr, who has been a mentor to me since the beginning. Without Glenn I would have never become a researcher and I am so grateful that we have become true colleagues over time.

I would not have been able to complete this PhD if not for the tremendous support I received in my work environment. I want thank Gary Newton, Physician-in-Chief at Mount Sinai Hospital, who encouraged me to take my time and not rush through the process, and Mount Sinai Hospital Department of Medicine for providing financial support for my studies. I also wish to thank my current and former Chairs of the Department of Medicine at the University of Toronto, Gillian Hawker and Wendy Levinson, for helping to protect my time for research and writing and for providing so much mentorship and support.

I am incredibly fortunate to be part of a research community that is so open, collegial and generous. I want to thank all of my brilliant colleagues and friends who were so generous with their time and support and were never too busy to answer a question (or ten). In particular I want to thank Geoff Norman for teaching me all about reliability and G-theory and Brian Hodges for encouraging me to go into a PhD with an open

mind and sincere curiosity. Completing a graduate program at a distance presents unique challenges so I was very fortunate to have the support of recent graduates Andrea Gingerich and Chris Watling to help me navigate the terrain. Finally, I want to express my deepest gratitude to my fellow scientists and all of the fellows, researchers and staff at the Wilson Centre for Research in Education, for inspiring and encouraging me every step of the way.

Dank je!